Supplementary Materials for "*Whole-genome haplotyping using a combination of dilution and statistical methods*" by Volodymyr Kuleshov[1,3]†, Dan Xie[2]†, Rui Chen[2]†, Dmitry Pushkarev[3], Zhihai Ma[2], Tim Blauwkamp[3], Michael Kertesz[3], Michael Snyder[2]*

Supplementary Figures 1-5

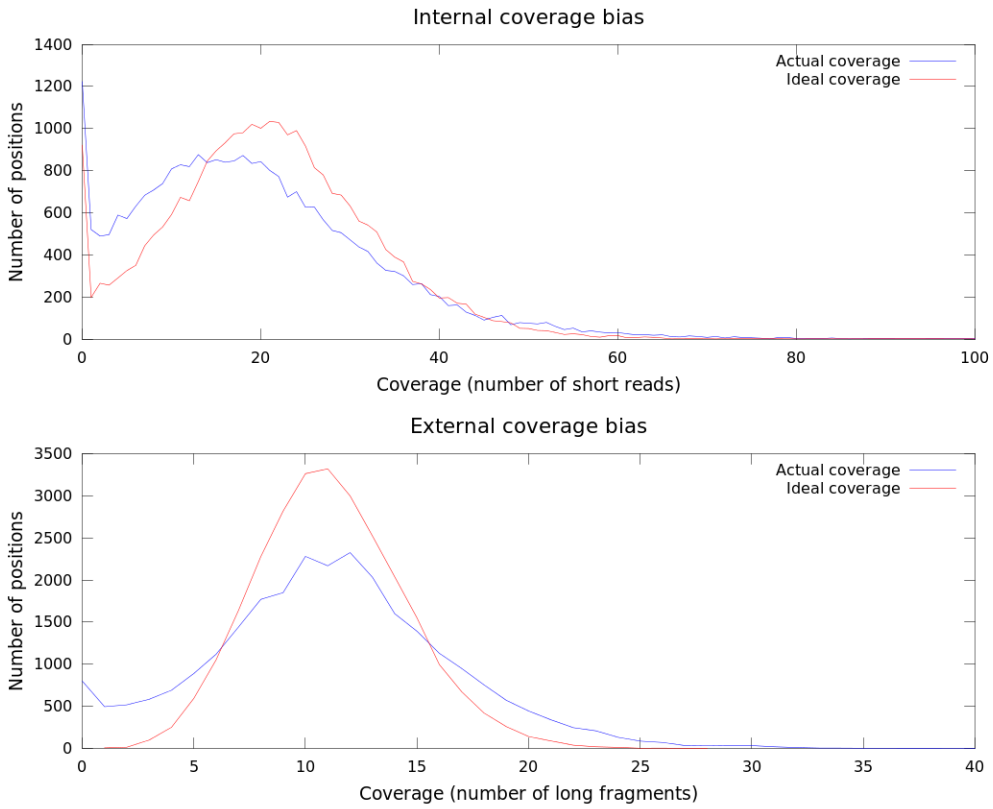
Supplementary Tables 1-15

Materials and Methods

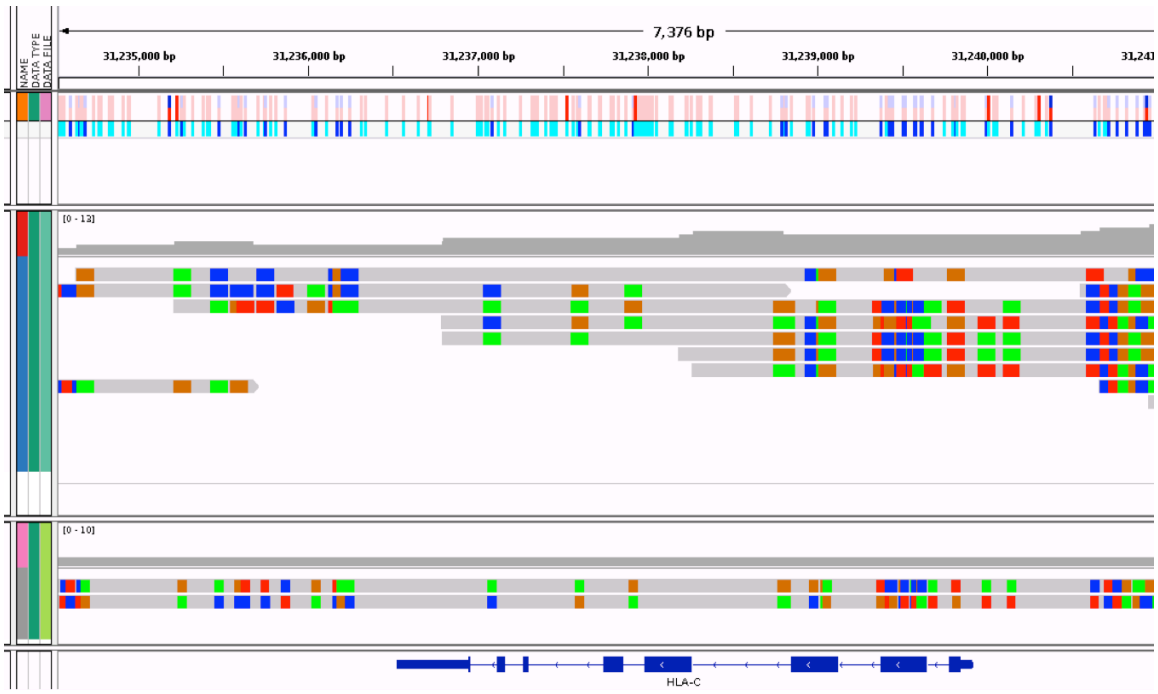
Description of the statistical phasing algorithm

Phased VCF files for the three subjects available online:
http://www.stanford.edu/~kuleshov/NA12878.vcf.gz
http://www.stanford.edu/~kuleshov/NA12891.vcf.gz
http://www.stanford.edu/~kuleshov/NA12892.vcf.gz

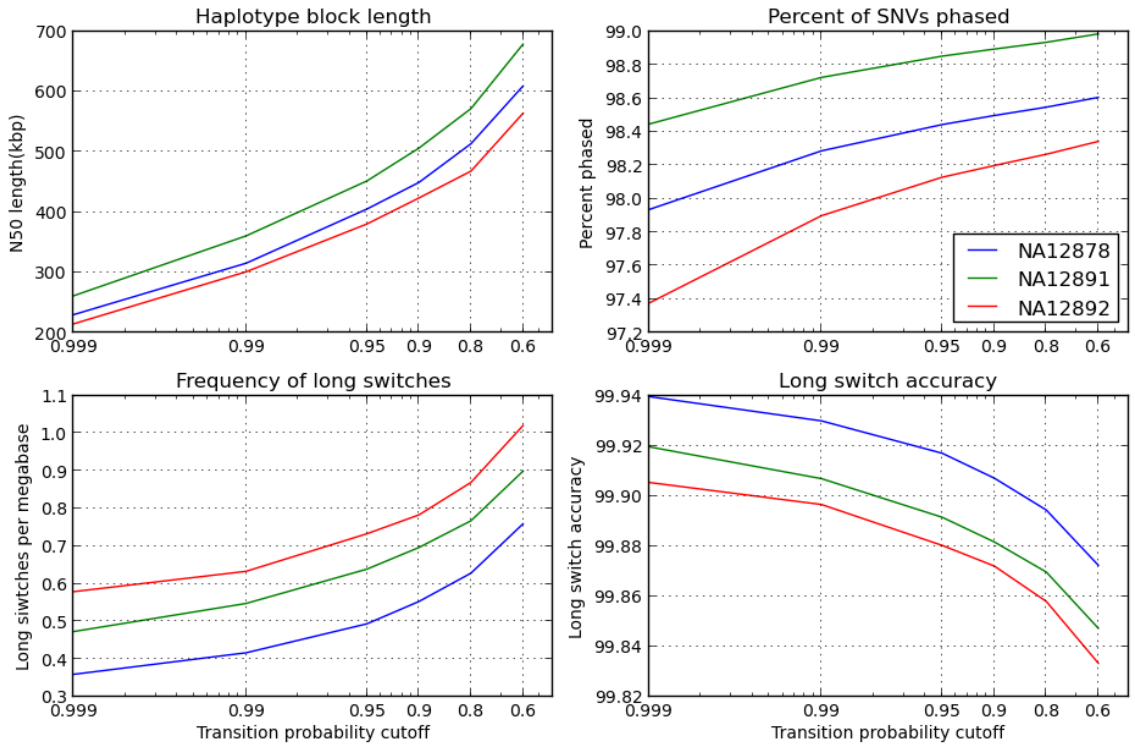
List of DMRs identified in sample NA12878



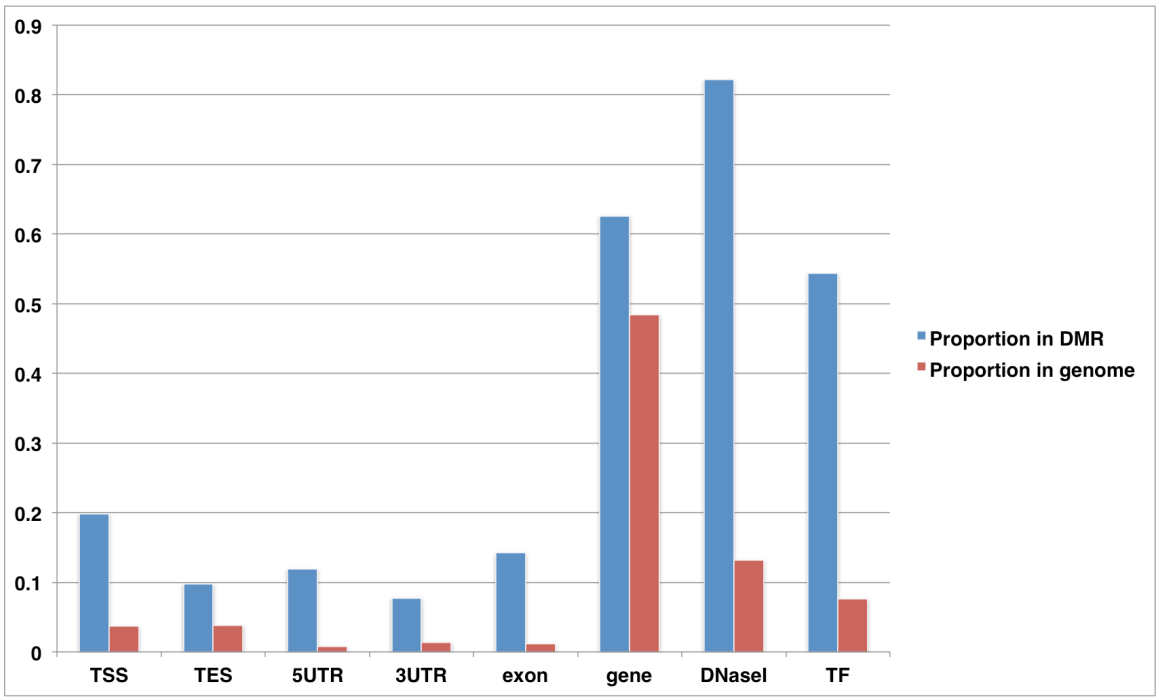
Supplementary Figure 1. Estimation of biases associated with the SLRH method. Both panels compare the empirical coverage histogram at heterozygous SNVs (blue curve) to one obtained from an idealized coverage distribution that exhibits no bias (red curve). The top panel shows the typical coverage by short reads; the ideal distribution was chosen to be free of internal biases. The bottom panel shows the typical coverage by long fragments; the ideal distribution exhibits no external biases.



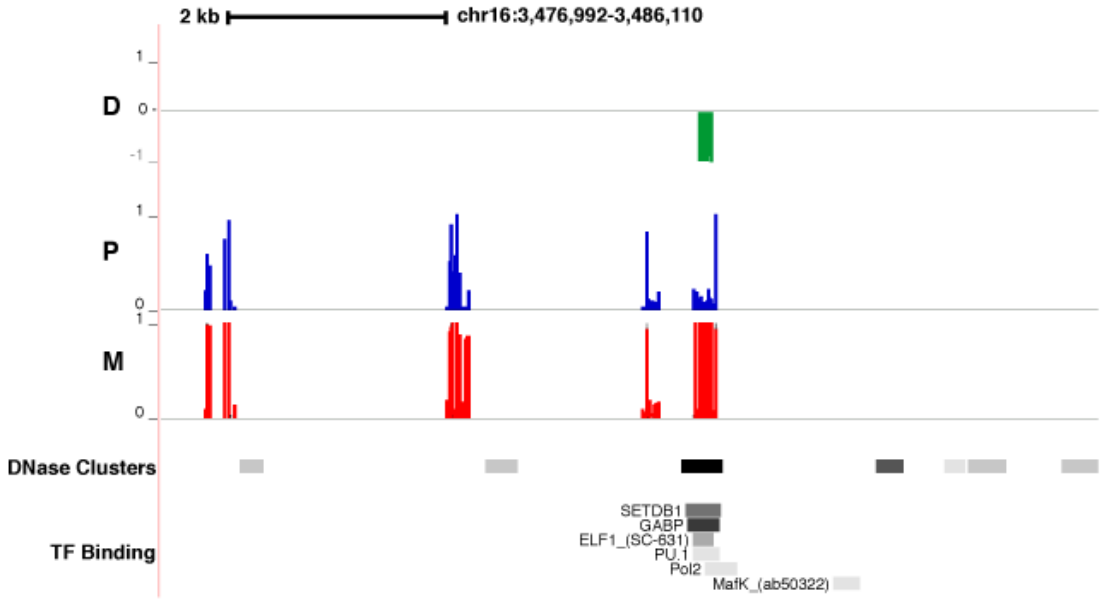
Supplementary Figure 2: Genome browser view of the phased HLA-C gene. Colored regions indicate heterozygous SNVs. Long fragments are connected at their overlapping heterozygous positions (middle track) into haplotype blocks (bottom track). Determining the phase of HLA genes has applications in host-donor matching for organ transplantation.



Supplementary Figure 3: Haplotyping results at different accuracy thresholds from a single 30 Gbp sequencing library. Long statistically constructed haplotype contigs are cut at positions where confidence scores are below a certain threshold, forming shorter but more accurate haplotype blocks. We evaluate the completeness (top panels) and the accuracy (bottom panels) of the smaller blocks at a series of error thresholds. The blocks are evaluated only over SNVs.



Supplementary Figure 4. Overlap of DMRs with different types of genomic regions. Blue bars indicate the proportions of DMRs overlapping with each type of genomic regions; red bars indicate the proportions of each type of genomics region in the human genome.



Supplementary Figure 5. Genome browser view of a typical DMR located at an intergenic region. The DMR resides on a DNase hypersensitive site and overlaps with many TF binding sites.

Supplementary Table 1: Comparison of SLRH to existing dilution haplotyping technologies. SLRH phases substantially more SNVs at a high level of accuracy than existing technologies while using substantially less sequencing data.

	SLRH	**Kaper et al.[2]**	**LFR[3]**	**Kitzman et al.**
Sample[1]	NA12878	NA12878	NA12877	NA20846
No. dilution aliquots used	384-768	192	384-768	116
Fragment size	8-9 Kbp (N50)	10-20 Kbp (average)	64-82 Kbp (N50)	~37 Kbp
Amplification method	Long-range PCR	MDA	MDA	Fosmid cloning
Sequencing library preparation method	Nextera v.2	Nextera v.2	CoRE and adapter ligation	Nextera v.1
NGS Platform	Illumina HiSeq	Illumina HiSeq	Complete Genomics	Illumina GAIIx
Library prep. time	2 days (6 hours hands on)	1 day	1 day	7 days
Bases sequenced (in addition to WGS)	30-60 Gbp	203-409 Gbp	238-496 Gbp	110 Gbp
% SNVs phased	99%	97%	92-97%	94%
N50 of haplotype blocks	450-560 Kbp	358 Kbp	530-600 Kbp	386 Kbp
Long switch accuracy[2]	99.90-99.92%	99.4%	2.2% of blocks containing switch errors[3]	n/a
Short switch accuracy	99.87-99.91%	99.7%	99.95-99.998%	99.7% concordance with (D'>0.9) HapMap SNPs
Phasing algorithm	Custom pipeline (*Prism*)	RefHap	Custom Complete Genomics pipeline	HapCut

[1] We report results for one representative sample. All the papers except Kitzman et al. sequenced

[2] In the Kaper et al. paper, "long switch accuracy" is referred to as "switch error rate"; "short switch accuracy is referred to as "accuracy considering switch errors".

[3] The long switch accuracy was not assessed within the LFR publication.

Supplementary Table 2: Sample statistics.

	NA12878	NA12891	NA12892
Genomic variants	3,760,133	3,742,625	3,787,927
SNVs	3,167,197	3,169,053	3,197,249
Indels	592,936	573,572	590,678
Heterozygous variants	2,249,071	2,220,244	2,305,926
Heterozygous SNVs	1,904,884	1,884,674	1,959,083
Heterozygous Indels	344,187	335,570	346,843

Supplementary Table 3: Library statistics before quality control filtering.

Library	Number of reads	Number of fragments	Total length of fragments	Coverage with fragments
NA12878-1	151,065,425	1,668,936	14,347,030,961	4.78
NA12878-2	152,045,561	1,694,672	14,445,637,123	4.82
NA12891-1	215,281,988	2,659,020	18,231,656,819	6.08
NA12891-2	224,946,933	3,037,427	22,047,999,725	7.35
NA12892-1	143,140,286	2,144,677	14,946,458,609	4.98
NA12892-2	191,480,706	2,714,314	21,319,831,141	7.11

Supplementary Table 4: Library statistics after quality control filtering. About 10-20% of fragments are discarded due to quality issues. The main reasons for discarding fragments are: insufficient internal coverage, a large number of positions with low q-scores, evidence of two different alleles at a heterozygous position in the same fragment, and a fragment length that falls within the top two percentiles.

Library	Number of fragments	Total length of fragments	Coverage with fragments	N50 fragment length
NA12878-1	1,415,940	12,216,607,477	4.07	8,925
NA12878-2	1,448,557	12,315,617,298	4.16	8,918
NA12891-1	2,181,780	15,029,824,324	5.02	7,950
NA12891-2	2,451,467	18,441,004,225	6.15	9,387
NA12892-1	1,771,595	12,417,265,735	4.14	7,987
NA12892-2	2,239,054	17,950,601,115	5.99	10,206

Supplementary Table 5: Phasing results for sample NA12878 at multiple thresholds measured over SNVs only. By adjusting the quality threshold, we can either obtain very long 1.1 Mbp haplotype contigs, or we can recover the extremely accurate blocks that do not involve statistical phasing. At both ends, blocks contain on average less than one long switch per Mbp.

Threshold	N50 (bp)	Percent phased	Variant N50 (variants)	Long switches per Mbp	Long switch accuracy	Short switch accuracy
0.5	1,106,509	99.23%	909	0.85	99.86%	99.87%
0.6	754,449	99.07%	657	0.64	99.89%	99.91%
0.7	684,268	99.05%	609	0.59	99.90%	99.91%
0.8	629,507	99.03%	565	0.54	99.91%	99.91%
0.9	563,801	99.00%	505	0.47	99.92%	99.91%
0.95	498,265	98.96%	455	0.41	99.93%	99.91%
0.99	393,510	98.85%	368	0.31	99.95%	99.92%
0.999	292,817	98.63%	280	0.27	99.96%	99.92%
0.9999	134,142	96.90%	153	0.26	99.96%	99.93%

Supplementary Table 6: Phasing results for sample NA12878 at multiple thresholds measured over both SNVs and indels. By adjusting the quality threshold, we can either obtain very long 1.1 Mbp haplotype contigs, or we can recover the extremely accurate blocks that do not involve statistical phasing. At both ends, blocks contain on average less than one long switch per Mbp.

Threshold	N50 (bp)	Percent phased	Variant N50 (variants)	Long switches per Mbp	Long switch accuracy	Short switch accuracy
0.5	1,100,074	94.30%	1006	1.02	99.85%	99.71%
0.6	750,102	94.16%	738	0.80	99.89%	99.74%
0.7	683,003	94.15%	685	0.74	99.89%	99.74%
0.8	628,233	94.13%	634	0.68	99.90%	99.75%
0.9	560,936	94.10%	566	0.62	99.91%	99.75%
0.95	496,623	94.07%	511	0.56	99.92%	99.75%
0.99	391,275	93.99%	412	0.45	99.94%	99.75%
0.999	290,660	93.81%	315	0.41	99.94%	99.76%
0.9999	132,146	92.43%	171	0.44	99.94%	99.76%

Supplementary Table 7: Overview of performance at the 0.9 accuracy threshold over SNVs only. We evaluate haplotype blocks obtained by introducing cuts in the raw statistically assembled haplotype contigs whenever a confidence score is below 0.9.

	NA12878	NA12891	NA12892
Fragment N50 (bp)	8,922	8,294	8,637
Number of fragments	2,868,739	4,633,247	4,004,932
Number of local blocks	304,023	216,816	294,805
Percent phased locally	93.43%	97.02%	94.16%
N50 of local blocks (bp)	65,843	80,388	71,359
N50 of global blocks (bp)	563,801	647,599	578,217
Percent phased globally	99.00%	99.25%	98.89%
Long switches per Mbp	0.47	0.68	0.81
Long switch accuracy	99.92%	99.89%	99.87%
Short switch accuracy	99.91%	99.84%	99.79%
Absolute accuracy	95.95%	93.05%	93.50%

Supplementary Table 8: Overview of performance at the 0.9 accuracy threshold over both SNVs and indels. We evaluate haplotype blocks obtained by introducing cuts in the raw statistically assembled haplotype contigs whenever a confidence score is below 0.9. Compared to results taken over SNVs only, the biggest change is an increase in short switching events.

	NA12878	NA12891	NA12892
Fragment N50 (bp)	8,922	8,294	8,637
Number of fragments	2,868,739	4,633,247	4,004,932
Number of local blocks	304,023	216,816	294,805
Percent phased locally	88.59%	92.19%	89.43%
N50 of local blocks (bp)	65,843	80,388	71,359
N50 of global blocks (bp)	560,936	643,067	574,192
Percent phased globally	94.10%	94.61%	94.09%
Long switches per Mbp	0.62	0.81	0.99
Long switch accuracy	99.91%	99.88%	99.86%
Short switch accuracy	99.75%	99.71%	99.64%
Absolute accuracy	95.80%	92.93%	93.33%

Supplementary Table 9: Percent of genes and novel variants phased. All novel variants were phased at the local stage, therefore we cannot improve phasing performance by increasing the accuracy threshold as we did above.

	Novel variants phased	Novel variant switch accuracy	Percent of genes phased
NA12878	73.10%	96.25%	89.78%
NA12891	75.86%	97.75%	88.86%
NA12892	73.16%	98.16%	88.47%

Supplementary Table 10: Phasing performance over SNVs from 30 Gbp of sequencing (0.9 accuracy threshold). We ran the bioinformatics pipeline independently on a single phasing library for each sample. The resulting blocks were almost as accurate and about 100 Kbp shorter than when two phasing libraries were used.

	NA12878 (lib. 1)	NA12891 (lib. 1)	NA12892 (lib. 1)
Fragment N50	8,925	7,950	7,987
Number of fragments	1,415,940	2,187,579	1,771,595
Number of local blocks	459,759	306,137	487,400
Percent phased locally	86.79%	93.37%	86.33%
N50 of local blocks	37,521	43,516	31,174
N50 of global blocks	449,802	506,575	423,854
Percent phased globally	98.54%	98.89%	98.20%
Long switches per mega base	0.57	0.70	0.78
Long switch accuracy	99.90%	99.88%	99.87%
Short switch accuracy	99.88%	99.79%	99.78%
Absolute accuracy	95.70%	93.51%	93.42%

Supplementary Table 11: Phasing performance over both SNVs and indels using 30 Gbp of sequencing (0.9 accuracy threshold). We ran the bioinformatics pipeline independently on a single phasing library for each sample. The resulting blocks were almost as accurate and about 100 Kbp shorter than when two phasing libraries were used.

	NA12878 (lib. 1)	NA12891 (lib. 1)	NA12892 (lib. 1)
Fragment N50	8,925	7,950	7,984
Number of fragments	1,415,940	2,187,579	1,778,945
Number of local blocks	459,759	306,137	487,400
Percent phased locally	82.47%	89.12%	82.34%
N50 of local blocks	37,521	43,516	31,174
N50 of global blocks	448,240	504,180	420,869
Percent phased globally	93.63%	94.53%	93.48%
Long switches per mega base	0.79	0.77	1.03
Long switch accuracy	99.89%	99.89%	99.86%
Short switch accuracy	99.61%	99.55%	99.50%
Absolute accuracy	95.45%	93.39%	93.13%

Supplementary Table 12: Evaluation of two replicate libraries for sample NA12878 using 30 Gbp of sequencing. The two replicates are highly concordant and exhibit a small loss in performance compared to when two libraries are used.

	NA12878 (#1 + #2)	NA12878 (#1)	NA12878 (#2)
Fragment N50	8,922	8,925	8,918
Number of fragments	2,868,739	1,415,940	1,448,557
Number of local blocks	304,023	459,759	484,632
Percent phased locally	93.43%	86.79%	85.69%
N50 of local blocks	65,843	37,521	37,902
N50 of global blocks	563,801	449,802	449,627
Percent phased globally	99.00%	98.54%	98.50%
Long switches per mega base	0.47	0.57	0.55
Long switch accuracy	99.92%	99.90%	99.91%
Short switch accuracy	99.91%	99.88%	99.87%

Supplementary Table 13: Statistical phasing accuracy for sample NA12878 at a confidence threshold of 0.9. Switch events were separated into three categories, depending on whether they occurred between two locally phased blocks of two SNVs or more, between two isolated SNVs, or between a block and an isolated SNV. Accuracies were calculated between adjacent blocks phased relative to each other with a confidence score of 0.9 or more.

	Accuracy
Switch accuracy between two local blocks	97.3%
Switch accuracy between a local block and an isolated SNV	98.4%
Switch accuracy between two adjacent isolated SNVs	99.4%
Total statistical switch accuracy	98.8%

Supplementary Table 14: Phasing performance on a simulated phasing library (equivalent to 30Gbp of input reads) with no sequencing or PCR errors. Results are presented at a 0.9 accuracy threshold over SNVs only.

	NA12878
Fragment N50 (bp)	7,000 bp
Percent phased locally	97.94%
N50 of local blocks (bp)	61,578 bp
N50 of global blocks (bp)	496,240 bp
Percent phased globally	99.38%
Long switches per Mbp	0.64
Long switch accuracy	99.89%
Short switch accuracy	99.98%
Absolute accuracy	95.65%

Supplementary Table 15: Phasing performance on a simulated phasing library (equivalent to 30Gbp of input reads) with no sequencing or PCR errors. Results are presented at a 0.9 accuracy threshold over both SNVs and indels.

	NA12878
Fragment N50 (bp)	7,000 bp
Percent phased locally	96.23%
N50 of local blocks (bp)	61,578 bp
N50 of global blocks (bp)	495,310 bp
Percent phased globally	97.59%
Long switches per Mbp	0.72
Long switch accuracy	99.90%
Short switch accuracy	99.98%
Absolute accuracy	95.64%

Materials and Methods:

Preparation of a Phasing Library

To prepare a phasing library, 1 microgram of genomic DNA was sheared using a Covaris g-Tube (3,200xg for 2x1 min). The 8-10Kb DNA fragment range was isolated from a 0.8% Clonewell E-gel (Life Technologies, Grand Island, NY, USA) using the QIAquick Gel Purification Kit (Qiagen). Isolated DNA fragment ends were blunted, 5'-phosphorylated, A-tailed, and ligated to dT-tailed adapters (see below) using the NEBNext Quick DNA Library Prep Master Mix Set for 454 (New England Biolabs), following the manufacturer's instructions.

Adapter (Forward): 5'-
CATCTCATCCCTGCGTGTCTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGT
ACGCTTGCAT-3'

Adapter (Reverse): 5'-(Phos)-
TGCAAGCGTACTGTCTCTTATACACATCTGACGCTGCCGACGAGACACGCAG
GGATGAGATGG-3'

Excess adapters and enzymes were removed using AMPure XP SPRI beads (Beckman Coulter Genomics, Danvers, MA, USA), and the concentration of amplifiable 8-10kb DNA fragments was determined by qPCR relative to 10Kb fragments of known concentration.

The 8-10Kb fragments were then diluted to 3000-6000 amplifiable molecules per well of a 384-well plate, and PCR-amplified for 13-15 cycles (94°C 15 sec, 65°C for 9 min) using adapter-specific primers (5'-CCATCTCATCCCTGCGTGTCTCG-3') and LongAmp polymerase (New England Biolabs). The average number of molecules within each well was kept around 3000-6000 to reduce the complexity of unique DNA sequences, which is important to aid fragment calling and avoid cross-phasing of fragments. Each resulting pool of amplified molecules was Tagmented using the Nextera DNA transposase (Illumina), end-repaired (72°C for 4 min), and sequencing adapters (Nextera Index Kit, Illumina) with barcodes unique to each well were incorporated through limited cycle PCR (6 cycles of 94°C 15 sec, 65°C for 4 min). The resulting sub-libraries were pooled together, purified using the QIAquick PCR Purification Kit (Qiagen), size selected by excising the 400-800 bp fragments from 2% SYBR Safe E-gels (Life Technologies) using the QIAquick Gel Extraction Kit (Qiagen), further amplified using PhusionGC polymerase (New England Biolabs) with Primer1 and Primer2 from the Nextera DNA Sample Prep Kit (Illumina®-Compatible) (EpiCentre), and purified using the Zymo Clean and Concentrate Kit-5 (Zymo). Sequencing libraries were quantitated using by qPCR (KAPA) and sequenced on Illumina HiSeq2000 sequencers using a 2x100bp plus single 8bp index read recipe.

DNA for HapMap samples NA12878, NA12891, NA12891 was obtained from lymphoblastoid cell lines (GM12878, GM12891, GM12892) available from the Coriell Institute for Medical Research.

The samples were whole-genome sequenced to a depth of 50X on an Illumina HiSeq 2000 instrument (Supp. Table 2) as part of the Illumina Platinum Genomes Project[4]. These requirements are comparable to those of earlier publications[1-3].

Comparison to the method of Voskoboynik et al.[5]

The above protocol adapts the LR-Seq technology that has been recently used to assemble the genome of B. schlosseri[5]. Our method has some differences to LR-Seq. LR-Seq, about 300 fragments wells are placed per well, compared to 5000 for SLRH. The individual fragments in each well are sequenced to a high depth in LR-Seq (50X, in order to perform de-novo subassembly); we sequence the fragments to a depth of 1-2X.

Additionally, the phasing protocol has been streamlined relative to LR-Seq in several ways. The most important modifications include: the elimination of two intermediary DNA purification steps; using the Nextera v.2 library preparation protocol; the addition of end-markers at the end of fragments; performing the initial fragmentation in a G-Tube, as opposed to using a HydroShear.

Assessment of PCR bias

To assess biases introduced by PCR, we compared the empirical read coverage at heterozygous SNVs to one obtained from sampling a uniform coverage distribution. Although we observe biases, they only affect a small number of SNVs.

We focus our attention on two types of coverage: internal and external. Internal coverage refers to the distribution of short reads within a typical long fragment. External coverage refers to the distribution of long fragments across the whole genome. Unevenness in each type of coverage is accordingly referred to as internal or external bias.

In Supplementary Figure 5, we plot the histogram of the internal and external coverages at heterozygous SNVs on chromosome 22. We derive these plots from data from the two phasing libraries of HapMap sample NA12878.

The top panel compares the empirical coverage histogram at heterozygous SNVs (blue curve) to one obtained from an idealized coverage distribution that exhibits no internal bias (red curve). The ideal coverage c_j (red curve) at a heterozygous position j is the sum $\sum_i c_{ji}$ of the coverages c_{ji} within each long fragment i that spans j; the c_{ji} were obtained by simulating a uniform internal coverage within long fragment i. We carried out this simulation by sampling each c_{ji} independently from a binomial distribution with parameters $n = 2.03 * l_i/101$ and $p = 101/l_i$, where l_i is the length of fragment i, 101 is the length of a short read, and 2.03 is the empirically derived average coverage of long fragments by short reads. In other words, the ideal curve assumes that short reads are distributed uniformly within each long fragment, but the long fragments are distributed across the genome in the same way as in the empirical data; this isolates the effects of internal bias.

The bottom panel compares the histogram of the empirical coverage by long fragments at heterozygous SNVs (blue curve) to an idealized distribution of long fragments that exhibits no external biases. The blue curve represents the typical number of long fragments that span a heterozygous position; a heterozygous SNV is said to be spanned by a long fragment if it falls between the start and end position of that fragment. The ideal coverage curve was derived under the assumption that long reads are distributed uniformly at random across the genome. We modeled this assumption by sampling a coverage c_j at every heterozygous position j from a binomial distribution with parameters $n = 11.12 * 5000/L$ and $p = 5000/L$, where 5000 was taken to be the length of a typical long fragment, L was the length of the reference sequence for chromosome 22, and 11.12 was the coverage of the chromosome by long fragments observed within the real data.

In the top panel, we observe more mass on the left of the blue histogram relative to the red histogram, indicating that many positions have less coverage than one would expect in the ideal case. However, there are only 302 more uncovered positions in the real data than in the ideal data (1.0% of the heterozygous positions on chromosome 22); thus even at a low internal coverage of 2x, relatively few positions are impossible to phase due to internal bias problems.

In the bottom panel, 800 heterozygous positions (2.9% of all positions) cannot be phased because of external bias; in the ideal case, there are no uncovered positions. This difference is larger than in the top panel, suggesting that the biases inherent to our method cause certain fragments not to amplify at all in certain regions of the genome, as opposed to some parts of a fragment not being amplified.

However, the above statistics suggest that only 4% fewer SNVs (relative to the ideal setting) cannot be phased by a purely molecular approach due to various biases. Interestingly, our local phasing rate of 93.4% (Supplementary Table 7) comes close to this upper bound of 96%.

Finally, we tested the impact of sequencing and PCR errors on the accuracy of our assembly by generating another simulated dataset for sample NA12878. We sampled 7Kbp-long fragments uniformly at random from the trio-phased genome of NA12878 (with each fragment coming from a single chromosome), such as to cover the genome at a depth of 6X (roughly the equivalent of one 30Gbp phasing library). Each long fragment was sampled with short reads at an internal depth of 2X, and each short read was generated from the trio-phased VCF of NA12878 with an error probability of 10^{-5}. This dataset was meant to represent an ideal input containing essentially no sequencing or PCR errors.

We ran our phasing pipeline over this dataset; results are presented in Supplementary Tables 14-15. Overall, we noticed an improvement in phasing quality. Substantially more variants were phased at the local stage into longer blocks. The subsequent global stage was also more complete, with an N50 length longer by about 100 Kbp, and with about 0.4% more SNV phased.

However, the difference in quality of these haplotypes was less significant. Long switch accuracy was similar in both cases, and actually 0.02% lower on the simulated dataset (this is understandable, since the haplotype blocks were substantially longer and contained more variants). The greatest improvement was found in short switch accuracy, suggesting that sequencing errors and PCR artifacts mainly introduce point errors at individual variants without significantly affecting the long-range phase information.

Computational Data Analysis

Sequencing reads were aligned to the genome using the BWA aligner. Within each set sharing the same well-specific barcode adapter, the reads were clustered into groups separated from each other by at least 2 Kbp. With high probability, reads within the same groups belong to the same long fragment. See Supplementary Table 3 for a summary of libraries sequenced.

Within each fragment, genomic variants were determined at set of heterozygous positions derived from an existing list of variants in VCF format. VCF files were obtained from an earlier whole-genome sequencing of the samples to a depth of 50X [4] on an Illumina HiSeq 2000 instrument (Supp. Table 2).

Once heterozygous variants in each fragment were determined, fragments were passed through quality control (Supp. Table 4). The reasons for discarding fragments were: insufficient internal coverage, a large number of positions with low q-scores, evidence of two different alleles at a heterozygous position in the same fragment, and a fragment length that fell within the top two percentiles.

Genotypes within a called fragment were sometimes inconsistent with the input genotyping; this occurred in one of two scenarios. If a position exhibited sufficient evidence ($p<0.01$) for two alleles, the entire fragment was discarded. This was done to detect and remove collisions between fragments in the same well; it was found to substantially increase the switch accuracy. About 1% of fragments were discarded in this manner. In the second scenario, only one allele was present at every position and the fragment was used for phasing. However, inconsistent positions were marked as part of the final output. About 1.4% of positions per library were marked as inconsistent.

Local phasing

First, at the *local* stage, long fragments were connected at their heterozygous SNVs using a dynamic programming algorithm. Dynamic programming is a technique that consists in finding the solution by first solving a set of smaller subproblems and then combining their solutions. Alternative algorithms for local phasing include RefHap[6] and HapCut[7]; they could probably be used to replace our method.

Our dynamic programming algorithm takes the approach of solving m subproblems (where m is the number of positions to phase), where each subproblem k (with

$1 \leq k \leq m$) consists in finding the optimal haplotypes for our data, assuming that the data is truncated to the first k positions.

For $k = 1$, that task is obviously easy to perform. For $k \geq 1$, we need to make the crucial observation that the optimal haplotypes up to position k consist of some assignment to parental chromosomes of the long fragments that span position k (call that assignment A), as well as of the best haplotypes over positions $1, \ldots, k-1$ under the conditions that the fragments that span both k and $k-1$ are assigned consistently (i.e., to the same parental chromosome). Let B denote an assignment of long reads to chromosomes at position $k-1$ and note that if we store the optimal solution over $1, \ldots, k-1$ for every assignment B, then we can compute a solution for position k by simply enumerating all assignments B that are consistent with A. Because our depth is relatively low, there are typically about ten fragments that span position $k-1$, and we can enumerate all possible assignments of these reads efficiently.

After repeatedly solving the subproblem for every k, we arrive at the full solution when $k = m$.

More formally, our algorithm is based on repeatedly solving the dynamic programming recursion

$$M[k, h_k, \chi_k] = \max_{h_{k-1}} \max_{\chi_k \sim \chi_{k-1}} \left(\sum_{i:\, k \in i} \log P(i(k) | h_k(\chi_k(i))) + M[k-1, h_{k-1}, \chi_{k-1}] \right)$$

where $M[k, h_k, \chi_k]$ is the log-likelihood of the best haplotype blocks given that the data is truncated to heterozygous positions 1 to k, that the haplotype at position k is $h_k \in \{0|1, 1|0\}$, and that the fragments at position k are mapped to the two parental chromosomes indexed by $\{0,1\}$ using the function $\chi_k : \{i | k \in i\} \to \{0,1\}$. In the last definition, the index i denotes fragments, and the notation $k \in i$ means that fragment i covers position k. The value $P(i(k) | h_k(\chi_k(i))$ is the probability of observing the variant $i(k)$ located at heterozygous position k in fragment i given that the fragment came from the parental chromosome $\chi_k(i)$, whose true allele is therefore taken to be $h_k(\chi_k(i))$. This probability is directly derived from the sequencing reads' q-scores. The expression that is maximized equals to the score of a particular assignment of fragments to parental chromosomes and of haplotypes at these alleles. It is maximized over the two possible haplotypes at the previous position h_{k-1}, and over all possible assignments of fragments to chromosomes at the previous position χ_{k-1} that are "consistent" with the current assignment χ_k. We enforce "consistency", in the sense that if a fragment spans positions k and k-1, then both χ_{k-1} and χ_k must assign it to the same chromosome. This is denoted by $\chi_{k-1} \sim \chi_k$.

The algorithm computes the $M[k, h_k, \chi_k]$ for all heterozygous positions $k = 1, \ldots, m$, and stores the haplotypes $h_1, \ldots, h_k$ associated with each $M[k, h_k, \chi_k]$. The final solution $\max_{h_m} \max_{\chi_m} M[m, h_m, \chi_m]$ corresponds to the assignment that maximizes the log-likelihood of the data $\sum_{k=1}^{m} \sum_{i:k \in i} \log P(i(k) | h_k(\chi_k(i)))$. Here is a pseudocode definition:

- Compute $M[1, h_1, \chi_1]$ for all h_1, χ_1.
- For position $k \in \{2, \ldots, m\}$:
    - For haplotype $h_k \in \{0|1,1|0\}$:
        - For all assignments χ_k:
            - Compute $M[k, h_k, \chi_k]$ using the dynamic programming recursion.
- Return $\max_{h_m} \max_{\chi_m} M[m, h_m, \chi_m]$

The running time of the above algorithm is linear in the number of heterozygous positions, and exponential in the genomic coverage (more precisely, in our implementation, complexity grows on the order of $O(2^{coverage})$). Because the genome is typically covered to a depth of ~8X, the exponential running time factor is not a problem in practice. At positions where the coverage is extremely high by chance, we discard the least informative fragments without noticeable loss in performance.

Global phasing

The end results of the local stage are short and accurate haplotype blocks whose characteristics are summarized in Supplementary Tables 7, 8. Next, at the *global* stage, these local blocks are phased with respect to each other using a statistical phasing algorithm to form long haplotype contigs. Most blocks are assigned a phase at this stage, with the exception of a small number that are most often comprised of a single novel heterozygous variant.

We defer the detailed definition of the statistical phasing algorithm to a supplementary document. In brief, it extends the Li and Stephens[8] model used in statistical packages such as IMPUTE2[9] or SHAPE-IT[10] to accept prior local phasing information. It uses a reference panel of phased haplotypes and a genetic map of the genome. This data was obtained from the latest version of the IMPUTE2 statistical analysis package. Traditional statistical algorithms typically have low accuracy; our method leverages locally-derived haplotype information to greatly reduce the number of possible haplotypes in any region, and thus improves the phasing accuracy (Supplementary Table 13).

Although in this work we ran Prism on European samples, the program also handles subjects of other ancestries, as well as subjects of mixed origins. Prism handles admixed populations in the same way as IMPUTE2, i.e. it phases the subject across small overlapping genomic regions of 100Kbp-1.5Mbp (the exact size affects the results very little), and then merges the results of each region. Within a particular region, it selects a reference sub-panel of K individuals that best describe the subject. Therefore, if the subject has a different ancestral origin within a particular region, the method will select the set of haplotypes from the appropriate population. In the case of sample NA12878, we found that using a population of mixed ethnicities produced similar results to ones obtained from a strictly European panel with a slightly improved global phasing accuracy of 94.73% (up from 94.69%), suggesting that the sample was well described by reference panel members of European ancestry. In more admixed samples, we expect to see a drop

in statistical phasing accuracy when using only a single ethnicity in the panel; we therefore suggest using the default setting and choosing the closest samples of each ethnicity within the entire window. For a more thorough discussion of this topic, we refer the reader to the IMPUTE 2 publication[9].

The global phasing stage produces long haplotype contigs as well as confidence scores that indicate the likelihood of making a switch error between two local blocks. Depending on the application, haplotype blocks can be constructed from the long contigs by introducing breaks whenever the confidence score between two statistically phased blocks is too low. One can obtain an estimate for the minimum accuracy over a region by multiplying the confidence scores. If two haplotypes appear to be equally likely, Prism will pick one at random, but will assign that haplotype a low score. We found the number of positions with such low scores to be on the order of 2-3%.

Evaluation criteria

Performance was assessed using a series of metrics: the N50 length of phased blocks, the percent of variants phased, as well as long and short switch accuracies and associated rates.

The N50 length of a set of haplotype blocks is defined as the length n at which half of the total bases in all the blocks are in blocks of length n or longer. We defined the length of a haplotype block to be the number of bases between the first and the last heterozygous variant in the block. The percent of variants phased was defined as the number of heterozygous variants in haplotype blocks that contain at least two heterozygous variants in total.

Accuracy was assessed in terms of the concept of a switch. A switch is said to occur at a heterozygous variant j if the mother's and the father's variants are inverted with respect to heterozygous variant j-1. For example, a switch occurs at heterozygous variant j if the allele on haplotype 0 at j is known to come from the mother and the allele on haplotype 1 is known to come from the father, whereas at j-1 it's the opposite: the allele on haplotype 0 comes from the father and the allele on haplotype 1 comes from the mother.

We differentiate two types of switches. A long switch (also referred to as a switch event) happens when the mother's and the father's variants are inverted for more than one position (e.g. MMFF). A short switch is said to occur when the parental variants are inverted for a single position (e.g. MMFM).
In this study, we focus our analysis on long switches. Such errors are more important as they substantially alter the haplotypes within a region (e.g. within a gene). They are also the most common type of error produced by the statistical phasing that SLRH uses. Finally, they are much easier to measure, as false short switches can be caused by genotyping errors in trio-based phasing.

We assess errors using long switch accuracy, which is defined as one minus the number of long switch events divided by the number of locations where switches can be

measured. Short switch accuracy is defined as one minus the number of positions with short switches divided by the number of heterozygous genomic variants. The rate of long switch events was defined as S * (1 – A) / 3200, where A is the long switch accuracy on SNVs, S is the number of heterozygous SNVs in the subject, and 3200 is the approximate length of a human genome in Mbp. Finally, we defined the absolute accuracy of a block as maximum of (a) the number of variants truly coming from the father and (b) the number of variants truly coming from the mother, divided by the number of variants whose provenance could be assessed (i.e. heterozygous variants that are not heterozygous in both the father and in the mother). Thus, if a haplotype block has alleles 00011000, and the true parental alleles are 00000000 and 11111111, then the absolute accuracy for this block is 6/8. The absolute accuracy we report is taken over all the blocks.

We generated a list of genes from the UCSC known genes database by performing a set union of the genomic regions associated with all the transcripts of a particular gene. We considered a gene to be phased if it did not contain the start of a haplotype block.

Analysis of haplotyping performance over rare variants

The performance of SLRH over rare variants was noticeably lower than over common ones, both in terms of phasing rate and accuracy. To better understand the reason behind the drop in phasing rate, we examined the distance of each variant to its closest neighbor. We found that rare variants were typically located farther from their neighbors than common variants. Whereas, the average proximity across all SNVs was 684.9 bp in sample NA12878, the average proximity for novel SNVs was 3001.9bp.

Next, we tried to explain the drop in accuracy over rare variants. This drop was present at the local stage, and therefore was not due to statistical methods. After manually examining rare SNPs with switch errors, we concluded that the error increase was because these variants were often miscalled: we were able to find a significant number of so-called novel positions at which none of ten clouds would contain an SNV. Sometimes the genomic region around the novel SNP seemed to show evidence for another event, such as a copy number variation.

To investigate this more formally, we computed for each locally phased position j a quality score, defined as

$$\prod_{i:j\in i} P(f_{ij}|h_{ij}),$$

where f_{ij} is the allele found on fragment i at position j and h_{ij} is the true allele at position j on the haplotype from which fragment i was deemed to originate. To compute such q-scores, one needs to run the local phasing algorithm to determine the most likely haplotypes and provenances of the long fragments.

As an example, if the true haplotype in a region was determined to be 01010, and a long fragment i with alleles 111 was deemed to originate from the middle three positions, then the probability $P(f_{i3}|h_{i3}) = P(1|0)$, which is the probability of having made a sequencing error at that position.

Our local phasing algorithm assumes that all positions are truly heterozygous. Therefore, if a SNP call is a false positive, many fragments that cover that positions will have associated scores of $P(f_{ij}|h_{ij}) = P(0|1)$ (probability of observing the reference, given that the true allele should be a variant), and the total quality score will be low.

Thus the above score can be taken as a rough indicator of inconsistency between the long fragments, and the data reported in the VCF. We compared the average q-score at novel SNPs with a switch error to the average q-score at non-novel SNPs with a switch error, and found the former to be noticeably lower: 0.3635 versus 0.6523. The average q-score across all positions was 0.9328. Within each class, q-scores were either very close to one, or very close to zero. Among novel SNPs with switches, 63% were below 0.1, and 36% were above 0.9; among non-novel SNPs with switches, 33% were below 0.1, and 65% were above 0.9. This again suggests that novel SNPs were in general much less supported by long fragment data, and perhaps were not truly SNPs. This in turn offers an explanation for their lower accuracy.

Additional assessment of concordance between replicate libraries of NA12878

To better assess the concordance of the two replicates of the HapMap sample NA12878, we evaluated the switch accuracy of replicate one using replicate two as the ground truth set. The two samples displayed very few long switching events relative to each other, with the long switch accuracy being equal to 99.9% at the 0.9 confidence score cutoff. However, the short switch accuracy between the replicates was equal to 99.3%, which was lower than the true short switch accuracy. This suggests that the long-range phase of the two replicates is highly concordant, but particular individual positions exhibit differences in phase, possibly due to being uncovered by reads in one of the samples.

We further assessed concordance by counting the SNVs that were phased in both samples. We found that 98.6% of all SNVs phased in replicate one were also phased in replicate two, again suggesting good concordance.

Whole-Genome Bisulfite Sequencing (MethylC-Seq)

Briefly, 5 micrograms of genomic DNA was used for preparation of each Illumina library. The genomic DNA was mixed with 25 nanograms of unmethylated Lambda DNA (Promega, Madison, WI, USA) as bisulfite conversion control, and sheared with the Covaris S2 system (Covaris, Woburn, MA, USA) with the following settings: use frequency sweeping; Intensity 4; Duty Cycle 10%; Bursts per Second 200; for a total of 2 minutes. Fragmented DNA was then concentrated with QIAquick PCR Purification Kit (QIAGEN, Hilden, Germany). The ends of the concentrated DNA were first repaired with the Epicentre End-ItTM DNA End-Repair Kit (Epicentre/Illumina, Madison, WI, USA), and a deoxyadenosine was added to the 3'-end with Klenow 3'->5' exo- enzyme (New England Biolabs, Ipswich, MA, USA), and ligated with Illumina's Early Access Methylation Adapter Oligo (Catalog # ME-100-0010, Illumina, Hayward, CA, USA). The ligated libraries were size-selected for an average insert size of 300 bp by agarose gel excision and extraction, and underwent bisulfite conversion using the MethylEasyTM Xceed Rapid DNA Bisulphite Modification Kit (Human Genetic Signatures Pty Ltd, North Ryde, Australia). Bisulfite-converted, adaptor-ligated DNA was then amplified with the uracil-tolerating PfuTurbo Cx Hotstart DNA polymerase (Agilent Technologies, Santa Clara, CA, USA) using the following program: 95°C 2 min, 98°C 30 sec, 10 cycles of (98°C 15 sec, 60°C 30 sec, 72°C 4 min), 72°C 10 min. The final amplified libraries were further purified with the Agencourt AMPure XP SPRI beads (Beckman Coulter Genomics, Danvers, MA, USA), and subjected to 101b paired-end sequencing using Illumina's HiSeq 2000 Sequencer.

Identification of Allele-specific DNA Methylation at Base Resolution

To identify genome-wide allele-specific DNA methylation, we performed MethylC-seq experiments on the DNA extracted from GM12878 cell line. A total of 796 million (13x per strand) uniquely mapped non-identical reads were obtained. The bisulfite conversion rates were 99.82%. We used the heterozygous SNPs from the phasing data to determine the allele-specific DNA methylation and assigned 48 million reads (6% of total reads) to one of the two parental alleles. We determined the allele-specific DNA methylation (ASM) using Fisher's exact test on each of the cytosine that was covered by at least 5 reads on each allele. A total of 216,034 ASM cytosines passed the significance test. About half of ASM events (63.32%) were from direct disruption or forming of CpG dinucleotides, which confirmed the accuracy of our method since DNA methylation occurs primarily at CpG sites. A total of 100,834 ASM events were left after removing the CpG-disrupted ASM sites. DMRs are identified by merging allele-specific methylated cytosines on the same allele that are less than 1000 bps apart. Merged genomic regions that contain 5 or more allele-specific methylated cytosines are reported as DMRs. All genomic coordinates were based on the GRCh37 (hg19) reference genome annotation.

Supplementary References

1. Kaper, F. et al. Whole-genome haplotyping by dilution, amplification, and sequencing. *Proceedings of the National Academy of Sciences* **110**, 5552-5557 (2013).
2. Peters, B.A. et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190-195 (2012).
3. Kitzman, J.O. et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* **29**, 59-63 (2010).
4. Illumina, Inc. The Platinum Genomes Project. *illumina.com* at <http://www.illumina.com/platinumgenomes/>
5. Voskoboynik, A. et al. The genome sequence of the colonial chordate, Botryllus schlosseri. *Elife* **2**, e00569 (2013).
6. Duitama, J. et al. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res.* **40**, 2041-2053 (2012).
7. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153-9 (2008).
8. Li, N. & Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**, 2213-2233 (2003).
9. 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).
10. Delaneau, O., Zagury, J. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Meth* **10**, 5-6 (2012).

Supplementary Material: The global phasing component of Prism

The global phasing algorithm constructs extremely long contigs from the short local haplotype blocks produced at a local stage. It takes as input a phased panel of individuals from the Thousand Genomes project [5], and uses linkage disequilibrium (LD) patterns inferred statistically from this panel to assign the most likely phase to each locally pre-phased block. Whereas traditional statistical algorithms have relatively low accuracy [1], our method is able to leverage locally-derived haplotype information to greatly reduce the number of possible haplotypes in any region, and thus improve phasing accuracy.

Overview

At a high level, our algorithm extends the Li and Stephens [4] model to sequencing data that comes pre-phased using long reads; existing algorithms based on this model are designed for microarrays [3, 2] and do not handle such input.

The intuition behind our algorithm is the same as behind ones based on the Li and Stephens [4] model. Given locally-phased data for which we only partially know the phase, we look for a way of representing this data as an "imperfect mosaic" of reference haplotypes. In other words we try to represent the partially-phased data as a combination of a small number of segments, such that within each segment, there is a reference haplotype that describes the observed data well. The phase of each block is then chosen to be that which best matches that of the chosen reference haplotypes.

One can also describe our algorithm in terms of the following generative model. This model postulates that the locally phased blocks are derived from the reference panel through a transcription process in which at every position, the observed haplotypes are copied from a pair of reference haplotypes $r^{(0)}$, $r^{(1)}$. At the first position, two imaginary "cursors" are initialized at a random pair of haplotypes; these cursors begin moving from left to right, starting at the leftmost genomic position $j = 0$. At every position, the first and the second cursor read the allele of their haplotypes; these letters become the allele at the first and the second haplotypes at the corresponding position of the locally-derived block covering that position. At some positions, however, an error occurrs in the transcription of the allele. With high probability, the cursors stay on the same reference haplotype. However, with small probability, they occasionally jump to a different haplotype, at which point it is that haplotype that becomes transcribed. The end result, is that the locally phased blocks are composed of a patchwork of reference haplotypes that describe the blocks in some mall segment.

Formally, this generative process corresponds to a hidden Markov model (HMM), where the $r^{(0)}$, $r^{(1)}$ are the hidden states. We refer the interested reader to [4] for more details on this approach, and why it is effective at inferring haplotypes. Here, we restrict ourselves to formally presenting the HMM and how it used to compute the phase of the local blocks.

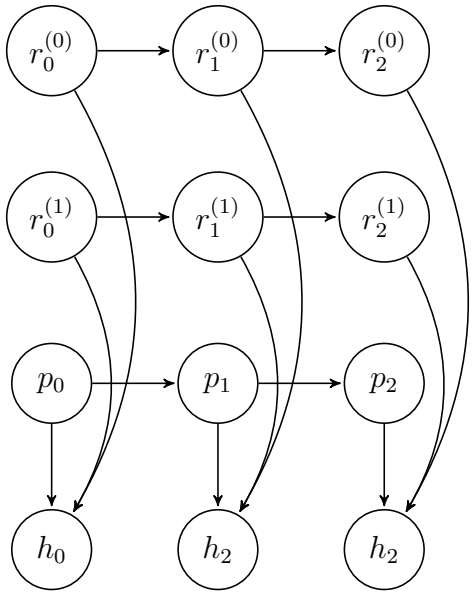
Notation

The global HMM is run on chromosome regions several hundreds of kilobases in length. Within these regions, we only consider positions that are variants either in the subject or in one of the reference panel samples.

We use m to denote the number of positions in the region, and denote each position by j. We use $0, 1$ to denote the two alleles at that position. The algorithm also assumes the existence of a panel of K phased haplotypes. We index panel haplotypes by $r = 1, ..., K$, and we denote by $r_j \in \{0, 1\}$ the allele of the member of the panel indexed by r. Our model assumes that data in this region is pre-phased using non-overlapping local blocks.

Graphical model

Our algorithm is based on the following hidden Markov model. Readers familiar with the subject will recognize it as an extension of the HMM of Li and Stephens [4].



The model has four types of variables: h, p, $r^{(0)}$, and $r^{(1)}$. There is an instance of each type of variable at every position $j = 1, ..., m$. The variable h_j corresponds to the haplotype of the subject at position j; it takes values in the set $\{0, 1\} \times \{0, 1\}$. Thus at heterozygous SNP positions, it is either $0|1$ or $1|0$; at homozygous SNPs it is always $1|1$, at reference positions it is always $0|0$. These values are determined at the local stage; thus if two adjacent heterozygous SNPs were found in local phasing to be on different chromosomes (e.g. two '1' alleles on different parental chromosomes), then their h-variables' values will be $0|1$, $1|0$, and not $0|1$, $0|1$. We denote the left and right alleles by $h_j(0)$, $h_j(1)$, respectively.

The $r_j^{(0)}$ and $r_j^{(1)}$ are indices of two reference panel haplotypes that describe the subject at position j. Each describe one chromosome of the subject. Finally, the p_j variable denotes the phase of the block that contains the SNP at position j; its value is in $\{0, 1\}$.

At every position, the emission probability equals

$$P(h_j|p_j, r_j^{(0)}, r_j^{(1)}) = \begin{cases} P(h_j(0)|r_j^{(0)})P(h_j(1)|r_j^{(1)}) & \text{if } p_j = 0 \\ P(h_j(0)|r_j^{(1)})P(h_j(1)|r_j^{(0)}) & \text{if } p_j = 1, \end{cases}$$

where each term $P(a|r_j^{(0)})$ is of the form

$$P(a|r_j^{(0)}) = \begin{cases} 1 - \lambda & \text{if } a = R(r_j^{(0)}) \\ \lambda & \text{if } a \neq R(r_j^{(0)}). \end{cases}$$

In the above equation, the variable λ is set as in the Li and Stephens model to $\lambda = \theta/(2(K + \theta))$, with $\theta = \left(\sum_{k=1}^{K-1} \frac{1}{k}\right)^{-1}$. Essentially, we observe the alleles of the reference panel haplotypes with a small probability of error. We refer the reader to [4] for the biological intuition behind the choice of these parameters.

The transition probabilities between the r-variables are also of the same form as in the Li and Stephens model:

$$P(r_j|r_{j-1}) = \begin{cases} \exp(-\rho_j/K) + (1 - \frac{\exp(-\rho_j/K)}{K}) & \text{if } r_j = r_{j-1} \\ \frac{\exp(-\rho_j/K)}{K} & \text{if } r_j \neq r_{j-1}. \end{cases}$$

The parameter ρ_j equals $4N_e d_j$, where N_e is the effective population size parameter (set to 15,000) in our algorithm, and d_j is the genetic distance between markers j and $j - 1$. This data is obtained from a genetic map that is an input to our algorithm. We again refer the reader to [4] for an explanation of the precise form of these equations, but the intuition is that at each step, the hidden state typically doesn't stage, except for a small probability of jumping to a different member of the reference panel.

The transition probabilities between adjacent p_j variables are uniform if SNPs j, $j - 1$ are in different blocks; otherwise, p_j and p_{j-1} are equal with probability one.

$$P(p_j|p_{j-1}) = \begin{cases} \frac{1}{2} & \text{if positions } j \text{ and } j_1 \text{ are in different locally phased blocks} \\ 1 & \text{if positions } j \text{ and } j_1 \text{ are in the same locally phased block and } p_j = p_{j-1} \\ 0 & \text{if positions } j \text{ and } j_1 \text{ are in the same locally phased block and } p_j \neq p_{j-1} \end{cases}$$

These transition probabilities enforce the locally-phased structure at the global phasing level.

Finally, the initial probabilities of the HMM are uniform.

Global phasing using the HMM

The above graphical model defines a probability distribution $P(\text{locally phased data}, p_{1:m}, r^{(0)}_{1:m}, r^{(1)}_{1:m})$. To infer the statistically most likely phase of the local blocks, we solve the following optimization problem:

$$\max_{p_{1:m}, r^{(0)}_{1:m}, r^{(1)}_{1:m}} P(\text{locally phased data}, p_{1:m}, r^{(0)}_{1:m}, r^{(1)}_{1:m}).$$

This problem can be solved efficiently using the Viterbi algorithm. Its solution is an optimal phase assignment to blocks $p^*_{1:m}$, and an optimal Viterbi path through the space of reference haplotypes $(r^{(0),*}_{1:m}, r^{(1),*}_{1:m})$.

In addition to finding the optimal phase, we also derive confidence scores from the forwards-backwards HMM probabilities using calculations that are typical for Hidden Markov Models. The most important score we use is the transition probability $P(p_j | p_{j-1}, h_{0:m})$. We use these scores to introduce breaks in our global haplotype contigs.

References

[1] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, sep 2011. 10.1038/nrg3054.

[2] Olivier Delaneau, Jean-Francois Zagury, and Jonathan Marchini. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1):5–6, dec 2012. 10.1038/nmeth.2307.

[3] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet*, 5(6):e1000529, jun 2009.

[4] Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233, 2003.

[5] Gil A McVean, David M Altshuler Co-Chair, Richard M Durbin Co-Chair, Gonçalo R. Abecasis, David R Bentley, Aravinda Chakravarti, Andrew G Clark, Peter Donnelly et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, oct 2012. 10.1038/nature11632.