SUPPLEMENTAL MATERIALS

Table of Contents:

1. Derivation of EBD^2 statistics:

In the manuscript we characterize EBD as a read depth pseudocount. By weighing read level data by base quality and mapping quality we are better able to guard against sequencing errors that may introduce false alternative alleles; a problem that is more acute in low coverage sequencing.

By inspiration, the squaring EBD stems from the pooled chi-squared statistic. We first establish the null hypothesis that each site is monomorphic; i.e. observations of alternative bases are due to sequencing or mapping error. For each site, *s*, we calculate the chi-squared statistic as follows (Eq.S1).

$$Pooled\ \chi_s{}^2 = \frac{\sum_{i=1}^{I}\left[a_{s,i} - e\left(a_{s,i} + r_{s,i}\right)\right]^2}{eT_s} \approx \frac{\sum_{i=1}^{I} a_{s,i}^2}{eT_s} \sim EBD^2$$

[Eq. S1]

In this equation *i* is indexed from *1 to I*, where *I* is the total number of sample. For each site *s*, the $a_i = \sum_k^K a_{i,k}\ and\ r_i = \sum_k^K r_{i,k}$ where k=1...K is the number of reads for alternative and reference nucleotides respectively. T_s is the total read depth over all samples for a given site. The numerator sums the variance over all samples. The denominator is the expectation under the null hypothesis. We can approximate this equation when $e\left(a_{s,i} + r_{s,i}\right)$ is close to zero, which is a reasonable assumption for low coverage sequencing. This simplifies to the squared read depth of the alternative divided by the expected read depth of sequencing error. Functionally, squaring the EBD value enhances the downward weighting of poor base and mapping quality, particularly in cases where the alternative allele occurs by sequencing error.

2. Derivation of site scoring statistics:

We derive the site-scoring statistic by taking inspiration from Fisher's 1954 test of heterogeneity which is based on the index of dispersion (Bennet and Hsu 1961). The index of dispersion is defined as the ratio of variance to the mean for a single set of observation within a trial. The test of heterogeneity, in contrast, computes the ratio of variances due to a series of different trials (Eq. S2). In our case, each sample BAM represents a single trial and the population represents the set of trials.

$$Fishers's\ \chi_s{}^2 - test\ of\ heterogeneity = \frac{\sum_{i=1}^{I}\left[a_{s,i} - e\left(a_{s,i} + r_{s,i}\right)\right]^2}{T_s e(1-e)}$$

[Eq. S2]

In our model, *e,* is the population probability of an alternative allele. The null hypothesis assumes that the population probability, *e,* is the same as the sample probability. Further this population probability is due to sequencing error as described in Section 1. In the test of heterogeneity the numerator represents the sum of variances for each sample. The denominator is the variance of the null hypothesis. Practically, this value of this statistic is increased if the site is truly polymorphic, i.e. there is extra-binomial variation.

Equation S2 performs well on real data, however we found that we could generate a more accurate statistic by utilizing a known property of bi-allelic SNPs. For a SNP, there are only three genotypes with corresponding own binomial parameters (0, 0.5, 1 for Ref/Ref, Ref/Alt, Alt/Alt respectively). The best fit statistic is achieved by utilizing the model that minimizes the value of Eq S3. We thus compose a goodness of fit test under the alternative hypothesis as:

$$\sum_{i=1}^{I} Min\left\{\left[a_{s,i} - 0\left(a_{s,i} + r_{s,i}\right)\right]^2, \left[a_{s,i} - \frac{1}{2}\left(a_{s,i} + r_{s,i}\right)\right]^2, \left[a_{s,i} - 1\left(a_{s,i} + r_{s,i}\right)\right]^2\right\}$$

[Eq. S3]

The Variance Ratio Statistic (Eq. S4) is a ratio of these two statistics (Eq. S2 and Eq. S3). The numerator represents the extra-binomial variation and the denominator acts as an approximate log likelihood where taking the minimum of the denominator maximizes the likelihood of the correct genotype. If there is high SNP variation in the denominator, then the statistic will be increased however if there is no SNP variation, then the signal will not be appreciably altered. If a site has a true alternative SNP, we would expect that it would have high extra binomial variation as measured by Eq. S2 and low SNP variation as measured by Eq. S3. We do not use the site level index, *s,* for readability.

$$Variance\ Ratio\ Statistic = \frac{\sum_{i=1}^{I}[a_i - e(a_i + r_i)]^2 - Te(1-e)}{\sum_{i=1}^{I} Min\left\{[a_i - 0(a_i + r_i)]^2, \left[a_i - \frac{1}{2}(a_i + r_i)\right]^2, [a_i - 1(a_i + r_i)]^2\right\}}$$

[Eq. S4]

We found that Eq. S4 outperforms the Fisher index of dispersion (Eq. S2) in sites where alternative bases are introduced by high mapping error (e.g. 10%) because it favors 'typical' genotype binomial ratios (0, 0.5 or 1) (data not shown)

3. SNP Site Detection

Using EBD and the variance ratio scoring statistic, we generated an unfiltered SNP site list that contained 34,656,295 candidate SNP sites on all autosome chromosomes with a threshold site score greater than 1.5.

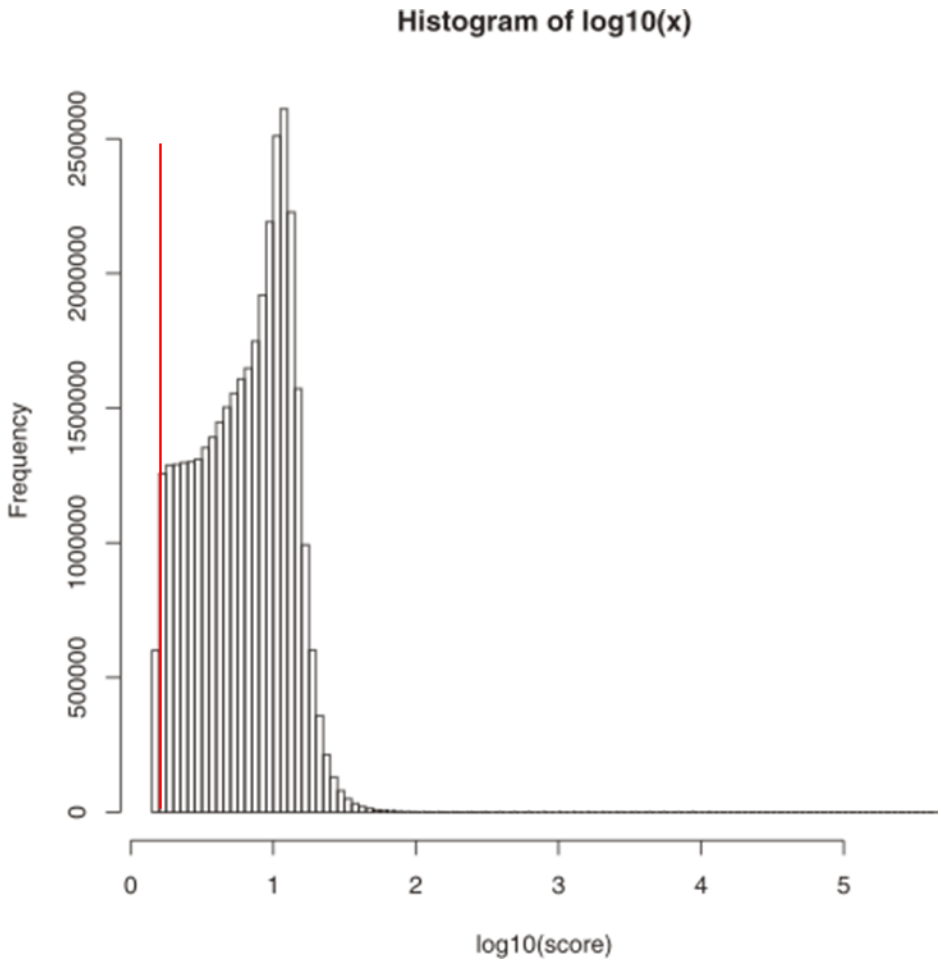

Figure S1: Distribution of Variance Site Scoring for unfiltered SNPs: We plot the distribution of site scores calculated using the variance site scoring method. In this figure the score cut-off is located at $\log_{10}(1.5) = 0.176$. Our unfiltered site list was composed of SNP sites with scores above 1.5.

SNP Site Discovery

In Table S3 we present the complete ***Unfiltered SNP site list*** with a total of 36,156,891 candidate SNP sites on autosomes with threshold site score threshold>1.5. We also include a ***Filtered SNP site list*** is generated by applying four filtering criterions (See below) on the unfiltered SNP site list. We obtain a total of 34,142,062 SNP sites on autosomes for the filtered list. This sites list was released in the 1000 Genomes interim release (Materials and Methods).

We found that the majority of the SNPs discovered are novel (78.3%) and that the Ti/Tv of known and novel SNPs are close to each other and highly correlated (r^2=0.97) across chromosome. The false positive rate is low in the sense of microarray (OMNI chip) validation: only 1954 sites out of 99,817non-polymorphic OMNI sites (1.9%) are falsely identified (Table S1).

SNP Site Filtering to increase specificity

Filtering of the site list is an optional procedure that increases the specificity of SNP discovery, although with some reduction in sensitivity. We employed four widely used heuristics: (1) maximum population read depth; (2) minimum population read depth; (3) strand bias as tested by 2x2 contingency table (rows are reference and alternative bases and columns are positive and negative strand), and (4) position bias. For tests involving population read depth, we removed SNP sites that deviated from the median values for the remaining SNPs. These criteria were fine-tuned on each sequencing platform to get remove SNP sites with an average Ti/Tv of ~1.4, a large deviation from genome wide expectations (Table S1). For strand and position bias, the mean of the reference base and the alternative base were evaluated by Fisher exact test. For each platform, we see progressive removal of low-quality SNPs and falsely discovered monozygotic OMNI sites.

		Sum EBD > 1.5 Median	Sum EBD < 0.2 Median	p(strand bias)<1e-3	p(position bias)<1e-3
Illumina	filtered	236,027	374,165	723,134	842,739
	Ti/Tv	1.34	1.35	1.25	1.29
	mono	150	64	436	603

		Sum EBD > 2.5 Median	p(strand bias)<1e-5	p(position bias)<1e-5
SOLiD	filtered	26,220	260,930	280,126
	Ti/Tv	1.23	1.55	1.56
	mono	17	53	183

		Sum EBD > 2 Median	Sum EBD < 0.05 Median	p(strand bias)<1e-3	p(position bias)<1e-3
454	filtered	38,577	238,971	2,958	16,169
	Ti/Tv	1.43	1.58	1.34	1.34
	mono	31	25	2	36

Table S1: Filtering Performance for each criterion by platform: Progressive application of filtering criteria resulted in removal of low-quality SNPs (low Ti/Tv) and removal of falsely discovered monomorphic OMNI sites.

After filtering, we removed approximately 5.5% of all sites; this included >30% more OMNI mono sites. This procedure also increased the overall Ti/Tv from 2.10 to 2.14. The filtered sites also had a Ti/Tv ratio ranging from 1.30-1.74, a large deviation from genome-wide expectations. Their Ti/Tv was not strongly correlated with that of unfiltered sites (r^2=0.36).

We calculate a pre and post filtering sensitivity and specificity by comparing discovery to OMNI based microarray.

	Sensitivity	Specificity
Before Filtering	97.9%	98.1%
After Filtering	93.6%	98.7%

Table S2: Pre and Post Filtering Sensitivity and Specificity. A total of 2,183,344 polymorphic and 99,817 monomorphic SNP sites were included in the OMNI microarray. Prior to filtering, SNPTools was able to correctly identify 2,138,395 sites, but falsely identified 1,946 sites. Post filtering, SNPTools correctly identified 2,043,844 sites but falsely identified 1,284 sites. Filtering increased the specificity by 0.6% but reduced sensitivity by 4.3%

Chr	Pre-Filtering Total	Ti/Tv	Removed SNPs	Removed SNPs Ti/Tv	Post-Filtering Total	Ti/Tv	Post-Filtering Known SNPs	Ti/Tv	Post-Filtering Novel (dbSNP 129)	% Novel	Ti/Tv
1	2,715,980	2.21	146,142	1.59	2,569,838	2.24	574,357	2.25	1,995,481	22.4	2.24
2	2,992,947	2.09	148,937	1.48	2,844,010	2.12	591,100	2.14	2,252,910	20.8	2.12
3	2,486,270	2.07	108,655	1.43	2,377,615	2.1	496,967	2.12	1,880,648	20.9	2.09
4	2,492,930	2.04	121,105	1.44	2,371,825	2.07	513,634	2.1	1,858,191	21.7	2.06
5	2,293,806	2.06	109,982	1.44	2,183,824	2.1	453,118	2.12	1,730,706	20.7	2.09
6	2,191,709	2.14	117,568	1.6	2,074,141	2.17	479,347	2.19	1,594,794	23.1	2.16
7	2,040,877	2.08	132,638	1.43	1,908,239	2.12	407,370	2.14	1,500,869	21.3	2.12
8	1,984,107	1.95	102,957	1.37	1,881,150	1.98	392,562	1.99	1,488,588	20.9	1.98
9	1,530,438	2	102,287	1.49	1,428,151	2.04	317,994	2.06	1,110,157	22.3	2.03
10	1,713,497	2.17	104,153	1.44	1,609,344	2.21	373,471	2.22	1,235,873	23.2	2.21
11	1,718,823	2.09	101,226	1.29	1,617,597	2.14	367,812	2.15	1,249,785	22.7	2.14
12	1,649,231	2.16	88,859	1.43	1,560,372	2.21	348,894	2.23	1,211,478	22.4	2.2
13	1,240,915	2.11	55,375	1.55	1,185,540	2.14	273,177	2.16	912,363	23	2.13
14	1,137,843	2.15	61,716	1.53	1,076,127	2.19	232,239	2.21	843,888	21.6	2.18
15	1,029,461	2.12	66,094	1.67	963,367	2.15	207,196	2.15	756,171	21.5	2.15
16	1,102,674	1.93	75,127	1.51	1,027,547	1.97	223,678	1.95	803,869	21.8	1.97
17	933,415	2.39	56,154	1.73	877,261	2.43	189,197	2.43	688,064	21.6	2.43
18	980,699	2.16	45,562	1.52	935,137	2.19	208,644	2.2	726,493	22.3	2.19
19	733,842	2.32	61,524	1.49	672,318	2.4	155,563	2.38	516,755	23.1	2.4
20	759,544	2.32	41,054	1.5	718,490	2.37	172,052	2.36	546,438	23.9	2.37

21	476,012	2.18	33,798	1.5	442,214	2.23	101,560	2.25	340,654	23	2.23
22	451,275	2.43	37,428	1.74	413,847	2.49	106,483	2.49	307,364	25.7	2.49
total	34,656,295	2.11	1,918,341	1.49	32,737,954	2.15	7,186,415	2.17	25,551,539	21.7	2.15

Supplement Table S3: Pre and Post filtering SNP site list for 1000 Genomes Phase 1 BCM call set.34,656,295 SNP sites were originally discovered with an average Ti/Tv ratio of 2.11. SNPs were filtered using 4 criteria (supplement) to produce a final list of 32, 737,954 SNPs with a Ti/Tv ratio of 2.15. 1,918,341sites (with Ti/Tv ratios ranging from 1.30-1.74) were removed from the unfiltered SNP list to form the filtered list. 21.7% of SNPs were previously discovered SNPs (dbSNP 129) and 78.3% of all sites were novel.

4. Expectation Maximization Algorithm for BBMM

We model each BAM as a flexible mixture of three binomials that represent each of the genotype classes. These genotype classes each have two variables, a weight coefficient w_v and a binomial probability p_v, In order to compute the value of these parameters we utilize the Expectation Maximization (EM) algorithm (Dempster et al. 1977). We begin by introducing a latent variable $z_{s,v}$ as a binary assignment variable which has a 1 of V representation $\sum_{v=rr,ra,aa} z_{s,v} = 1$ (Bishop, 2006). The probability of the latent variable at any site is then given by the weight coefficient, w_v such that $p(\boldsymbol{z_s}) = \prod_v \ {w_v}^{z_{s,v}}$. Similarly the observed data likelihood can be written as $p(r_s, a_s|\boldsymbol{z_s}) = \prod_v \ Binomial(r_s + a_s, p_v)^{z_{s,v}}$. The marginal distribution of the data is obtained by summing over the joint distribution all states of z, $\sum_z \ p(r_s, a_s|\boldsymbol{z_s})p(\boldsymbol{z_s})$. This simplifies into the sum of the weighted observed data likelihoods.

$$P(r_s, a_s) = \sum_{v=rr,ra,aa} w_v * Binomial(r_s + a_s, p_v)$$

Where

$$Binomial(r_s + a_s, p_v) = \binom{r_s + a_s}{a_s} {p_v}^{a_s}(1 - p_v)^{r_{s,}}.$$

To being the E-step, we assign values to the parameters${p_v}^0, {w_v}^0$. We then define the expectation the joint log likelihood with respect to the conditional distribution of $z_{s,v}$ with respect observed data and the previous parameters values as the following function O.

$$O(p, w|\ p^t, w^t) = \sum_s \sum_v \log\left((p\left(a_s, r_s\middle|z_{s,v}, p, w\right)\right)p(z_{s,v}|{p_v}^t, {w_v}^t, a_s, r_s)\,, t = 0,1,2..$$

For the M-step, we obtain an updated set of parameters $({p_v}^{t+1}, {w_v}^{t+1})$ by maximizing the above expectation.

$${p_v}^{t+1}, {w_v}^{t+1} = \max_{p,w} O(p, w|\ p^t, w^t)$$

We compute the updated parameters by taking the partial derivative of the function O, and solving for each of the parameters. These derivations can be found in Bishop (Bishop 2006)

.

5. Imputation and Phasing

Our imputation method derives from genetic coalescence based methods. However directly applying the Li and Stephens (Li and Stephens 2003) PAC method is computationally expensive – the estimation of multi-haplotype mosaic model by HMM is demanding due to the large number of hidden status ($O(I^2)$ states and $O(I^4)$ transitions). We improve computational time by restricting the number of population haplotypes that are used to create the mosaic. In our model, we employed a "constrained" template haplotype sampling scheme so that we can sample in constant time.

Pseudocode for Imputation and Phasing

```
Guess the haplotype of all samples randomly
Repeat b+m times, where b is the number of burn in cycles and m is the number of MCMC
iterations {
    For each individual i {
            Sample four haplotypes from the population randomly
            For w iterations {
                Change one of the four haplotypes
                    Accept the trial according to M-H criteria.
            } The remained haplotypes are the parental haplotypes for individual i
            Impute phased genotypes based on parental haplotypes.
    }
}
Repeat m cycles and output the haplotype probabilities and genotype likelihoods
```

To estimate haplotypes, we first initialize all haplotypes by randomly generating a possible haplotype given the set of observed GL for each individual. For each sample, we search for a set of 4 parental haplotypes $\boldsymbol{H_i^*}$ (also denoted as F0, F1, M0, M1) by proposing haplotypes from the population given observed genotype likelihoods for each sample and accepting or rejecting it according to a Metropolis Hastings acceptance criteria. We iterate this step, *w* times. Although the candidate space for

sampling is large$\sim O(I^4)$, by using the MH sampler to propose new parental haplotypes from the population of haplotypes, the sampler can evaluate the proposal in constant time $\sim O(I)$(Figure S2a). Once we have settled on a set of parental haplotypes, we can refine the sample's haplotype H_i, using a 4 state Hidden Markov Model, where the state of each site can be denoted as (F0, M0), (F0, M1), (F1, M0), (F1, M1), i.e. a mosaic combination of the four parental haplotypes $\boldsymbol{H_i}$*where the transition matrix is calculated from the recombination rate and genotype likelihoods and emission matrices is calculated from the mutation rate. We solve for the hidden states using the forward-backward algorithm (Bishop 2006). We iterate these steps, *m* times (Figure S2b) to produce accurate phased genotypes.

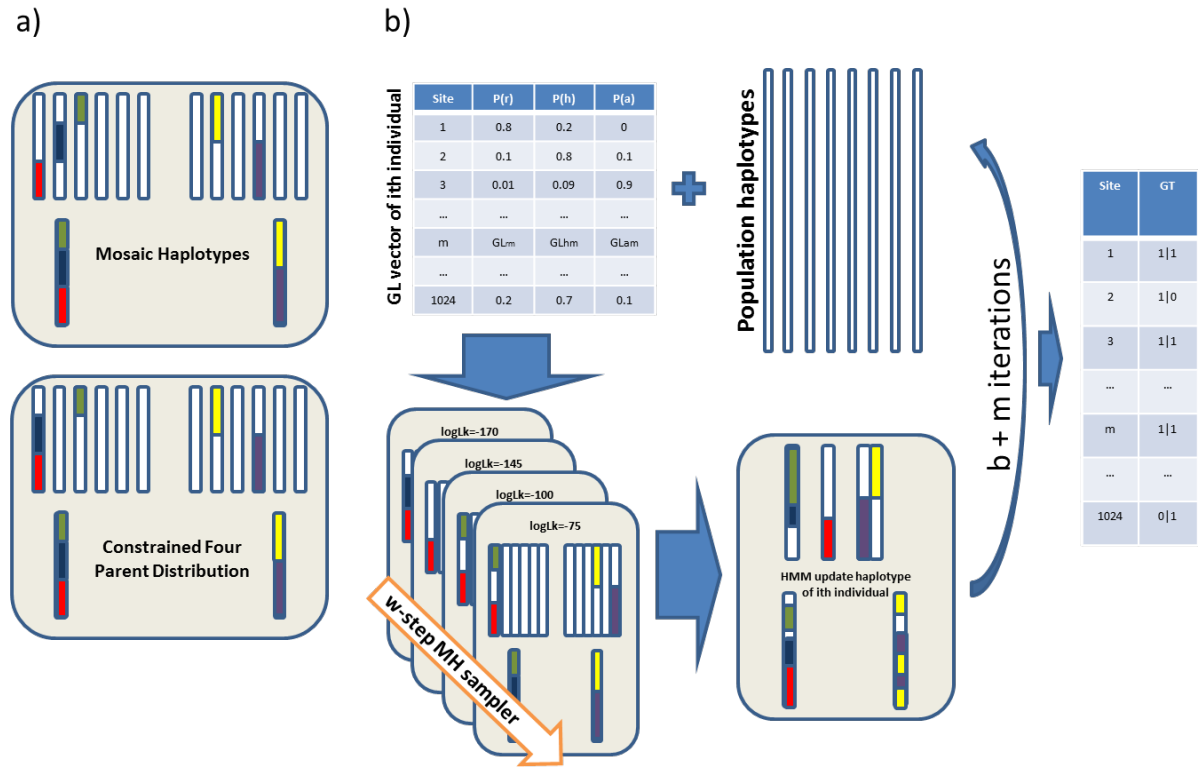

Figure S2: Imputation of phased genotypes using a constrained Li-Stephens method: Genotypes and haplotypes are imputed using a constrained mosaic haplotype model. (a). While mosaic haplotype model initializes each individual's haplotype as a mosaic of all haplotypes in the population, this is computationally expensive. Our model models the individuals as a mosaic of 4 parental haplotypes. This is more computationally tractable. (b) Given each individual's genotype likelihoods, 4 parental haplotypes are sampled from the population by a w-step Metropolis Hastings (MH) sampler. Once all parental haplotypes are established, genotypes and haplotypes for each individual are imputed given their parental haplotypes. This is repeated for *m* iterations until convergence.

Algorithm Development

The imputation and phasing process begins by dividing *S* sites into chunks of pre-set size. Smaller chunk sizes provide for more accurate modelling of local recombination rates but must be balanced against computational requirements. We select chunks of length 1024 in order to allow for parallel computation (See Section 6 below). For each chunk we setup a matrix of GLs for Ref/Ref, Ref/Alt, and Alt/Alt $g= (g_1...g_S)$ with S total variant sites (in the chunk) and a matrix $H= (H_1..., H_k,...,H_{2S})$ *of* haplotypes for each sample. To model recombination we create a vector $r=(r_1,...,r_{S-1})$ which is the probability of recombination, or "jump" probability between pairs of consecutive loci. Like Fearnhead and Donnelly, (Fearnhead and Donnelly 2001) we model this probability as $\frac{d_i*\rho}{2n+d_i*\rho}$ where d_i is the distance between marker *s* and *s+1* and ρ is the average recombination rate for the chunk.

To mimic the effects of recombination we model the loci as a Markov chain (1…S) with transition probability (T) between haplotypes, *k,* for individual *i* at locus *s*.

$P(H_{is} = k^*|H_{is} = k)$is equal to:

$$\begin{cases} (1-r_s)*(1-r_s)\ \ if\ k^* = k\ for\ both\ haplotypes \\ (1-r_s)*r_s\ if\ k^* = k\ for\ one\ haplotype\ and\ k^* \neq k\ for\ the\ other \\ r_s*r_s\ if\ k^* \neq k\ for\ both\ haplotypes \end{cases}$$

For each locus, *s*, we calculate four emission probabilities (*e*). These emission probabilities represent the four phased genotypes (0/0 –Ref/Ref, 1/0 –Ref/Alt, 0/1 –Ref/Alt, and 1/1 – Alt/Alt). The probabilities are calculated as the weighted sum of the product of a scaled mutation, μ, event for each genotype multiplied by the genotype likelihood for each genotype.

$$\begin{cases} P(0/0)_s = (1-\mu)*(1-\mu)*GL_s(Ref/Ref) + (1-\mu)*\mu*GL_s(Ref/Alt) + \mu*(1-\mu)*GL_s(Alt/Ref) + \mu*\mu*GL_s(Alt/Alt) \\ P(1/0)_s = (1-\mu)*(1-\mu)*GL_s(Alt/Ref) + (1-\mu)*\mu*GL_s(Ref/Ref) + \mu*(1-\mu)*GL_s(Alt/Alt) + \mu*\mu*GL_s(Ref/Alt) \\ P(0/1)_s = (1-\mu)*(1-\mu)*GL_s(Ref/Alt) + (1-\mu)*\mu*GL_s(Ref/Ref) + \mu*(1-\mu)*GL_s(Alt/Alt) + \mu*\mu*GL_s(Alt/Ref) \\ P(1/1)_s = (1-\mu)*(1-\mu)*GL_s(Alt/Alt) + (1-\mu)*\mu*GL_s(Alt/Ref) + \mu*(1-\mu)*GL_s(Ref/Alt) + \mu*\mu*GL_s(Ref/Ref) \end{cases}$$

The M-H Sampler:

In our "constrained Li- Stephens"(Li and Stephens 2003) algorithm, the sample haplotype is modelled as a mosaic combination of 4 parental haplotypes. To select these 4 parental haplotypes, we use an

M-H sampler to propose parental haplotypes by using the transition (*T*) and emission (*e*) matrices previously calculated. We initialize this process by randomly generating haplotypes for each individual and then randomly drawing four parental haplotypes, $\boldsymbol{H_i^*}$ from the population. We then calculate the likelihood of the current proposal $p(H_i|H_i^*,T,e)$. We then update one of the four parental haplotypes with a randomly drawn haplotype from the population. The proposed parental haplotypes are represented as H_i^{**}. We calculate the M-H acceptance probability as:

$$A(H_i^{**}, H_i^*) = \min\left[1, \frac{\mathrm{P}(H_i|H_i^{**}, T, e)}{\mathrm{P}(H_i|H_i^*, T, e)}\right]$$

We accept the probability $A(H_i^{**}, H_i^*)$:

- if $\mathrm{p}(H_i|H_i^{**}, T, e) > \mathrm{p}(H_i|H_i^*, T, e)$, we accept the proposed parental haplotypes, i.e. $H_i^* = H_i^{**}$
- if $\mathrm{p}(H_i|H_i^{**}, T, e) < p(H_i|H_i^*, T, e)$, we draw a random number from 0 to 1 and accept H_i^{**} if the random number is less than $\frac{\mathrm{p}(H_i|H_i^{**},T,e)}{\mathrm{p}(H_i|H_i^*,T,e)}$.
- Otherwise we keep the original four parental haplotypes, i.e. $H_i^* = H_i^*$

In order to ensure adequate mixing of the parental haplotypes, we use a *w*-step M-H sampler, where *w* = *number of folds* * *number of samples*. Increasing the number of folds increases the probability that all population haplotypes are sampled at least one time, during each M-H and HMM iteration.

The Hidden Markov Model (HMM):

Haplotypes exhibit cluster like patterns (Scheet and Stephens 2006). We can model the assumption that each allele will originate from a particular cluster with an HMM, where the transition (*T*) matrix can account for the assumption that nearby markers will likely arise from the same haplotype cluster (linkage disequilibrium) but that they will also be subject to recombination that scales with the distance between loci. In order to estimate the underlying haplotype states given the parental H_i* haplotypes, we utilize the forward backward algorithm (Bishop 2006)(Bishop 2006). As Baum's

forward backward algorithms have been discussed in depth by numerous authors (Browning and Browning 2007; Greenspan and Geiger 2004; Scheet and Stephens 2006; Bishop 2006; Li et al. 2011), we direct you to those references for more details.

Haplotype Merging:

The chunks are linked together to produce haplotypes by merging over the whole chromosome. All chunks are loaded into memory and the haplotypes are then merged using a 50kb sliding window by evaluating the phase position by position and finding the best match.

6. MCMC and Chunk Size for Genotype Imputation

Two key parameters affect the accuracy and speed of imputation: "chunk size" or number of sites to be imputed and the number of MCMC sampling cycles. We generated several genotype imputation call sets of chromosome 20 using different parameter settings and evaluated the results by comparing the genotype concordance against known genotypes from HapMap3, OMNI and Affymetrix Axiom data sets. We measure the error by the discordance rate (%) for the genotype classes, Alt/Alt, Ref/Alt and Ref/Ref and also evaluate an overall discordance rate and a non-ref/ref discordance rate. The best overall discordance rate was 0.61%, 0.67% and 0.62% when compared to HapMap3, OMNI and Axiom respectively was with MCMC =200, and chunk size =1024.

Reference	MCMC	Chunk Size	Ref/Ref	Ref/Alt	Alt/Alt	Total	Non-Ref
HapMap3	30	1024	0.36%	1.14%	1.44%	0.78%	1.67%
	50	1024	0.31%	1.04%	1.37%	0.71%	1.52%
	65	1024	0.29%	0.99%	1.32%	0.68%	1.45%
	100	1024	0.26%	0.94%	1.26%	0.64%	1.36%
	100	512	0.27%	0.92%	1.25%	0.63%	1.36%
	200	1024	0.25%	0.89%	1.21%	0.61%	1.30%
OMNI	30	1024	0.43%	1.67%	1.59%	0.82%	2.53%

	50	1024	0.40%	1.52%	1.53%	0.76%	2.34%
	65	1024	0.39%	1.44%	1.46%	0.73%	2.24%
	100	1024	0.37%	1.45%	1.43%	0.72%	2.21%
	100	512	0.37%	1.35%	1.40%	0.69%	2.14%
	200	1024	0.35%	1.34%	1.35%	0.67%	2.07%
Axiom	30	1024	0.27%	1.82%	1.55%	0.75%	2.25%
	50	1024	0.24%	1.69%	1.50%	0.70%	2.10%
	65	1024	0.23%	1.62%	1.45%	0.67%	2.01%
	100	1024	0.21%	1.58%	1.41%	0.65%	1.93%
	100	512	0.22%	1.53%	1.40%	0.64%	1.91%
	200	1024	0.21%	1.52%	1.37%	0.62%	1.86%

Table S4: Impact of MCMC cycles and chunk size on Chr20 Imputation Accuracy: (a) Discordance rates for imputed genotypes on chr20, created using different Chunk Sizes and MCMC, compared to genotypes from three array datasets, HapMap3, Illumina OMNI and Affymetrix Axiom. MCMC of 200 produced the lowest error rate.

While imputation accuracy can be improved with increased expenditure of computation resources, we found, as shown in Figure S2 for whole genome imputation, that after 1000 CPU equivalent months, that the improvement in error rate is marginal. A CPU month is a unit of computation defined as 1 CPU core of Intel Xeon E5520 working for 1 month.

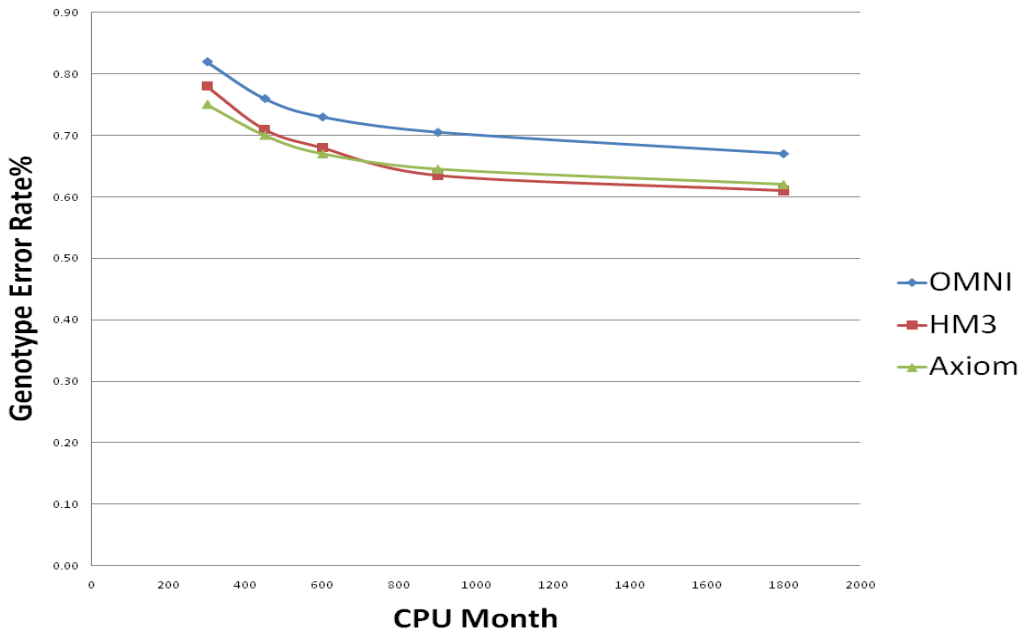

Figure S2: Computational burden of Imputation: We find that after 1000CPU months, increased MCMC cycles only result in marginal improvements in Genotype error.

References

Bennet BM, Hsu P. 1961. A sampling study of the power function of the binomial x2 “index of dispersion” test. *The Journal of hygiene* **59**: 449–55. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2134460&tool=pmcentrez&rendertype=abstract (Accessed November 23, 2011).

Bishop CM. 2006. *Pattern Recognition and Machine Learning*. Springer Science + Business Media, LLC, New York.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* **81**: 1084–97. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265661&tool=pmcentrez&rendertype=abstract (Accessed August 11, 2011).

Dempster AP, Laird NM, Rubin DB. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* **39**: 1–38.

Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. *Genetics* **159**: 1299–318. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461855&tool=pmcentrez&rendertype=abstract.

Greenspan G, Geiger D. 2004. Model-based inference of haplotype block variation. *Journal of computational biology : a journal of computational molecular cell biology* **11**: 493–504. http://www.ncbi.nlm.nih.gov/pubmed/15285904.

Li N, Stephens M. 2003. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **2233**: 2213–2233.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome research* **21**: 940–51. http://www.ncbi.nlm.nih.gov/pubmed/21460063 (Accessed July 18, 2011).

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics* **78**: 629–44. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1424677&tool=pmcentrez&rendertype=abstract.