Next generation genotype imputation service and methods

SUPPLEMENTARY NOTE

Sayantan Das[1,15], Lukas Forer[2,15], Sebastian Schönherr[2,15],
Carlo Sidore[1,3,4], Adam E. Locke[1], Alan Kwong[1], Scott I. Vrieze[5],
Emily Y. Chew[6], Shawn Levy[7], Matt McGue[8], David Schlessinger[9],
Dwight Stambolian[10], Po-Ru Loh[11,12] , William G. Iacono[8], Anand Swaroop[13],
Laura J. Scott[1], Francesco Cucca[3,4], Florian Kronenberg[2], Michael Boehnke[1],
Gonçalo R. Abecasis[1,16], and Christian Fuchsberger[1,2,14,16]

[1] Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
[2] Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria
[3] Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, Cagliari, Italy
[4] Università degli Studi di Sassari, Sassari, Italy
[5] Institute for Behavioral Genetics, University of Colorado, Boulder, CO, USA
[6] Clinical Trials Branch, Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health, Bethesda, MD, USA
[7] Hudson Alpha Institute for Biotechnology, Huntsville, AL, USA
[8] Department of Psychology, University of Minnesota, Minneapolis, MN, USA
[9] Laboratory of Genetics, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA
[10] Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA
[11] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
[12] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
[13] Neurobiology-Neurodegeneration and Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, MD, USA.
[14] Center for Biomedicine, European Academy of Bolzano/Bozen (EURAC), affiliated with the University of Lübeck, Bolzano, Italy.
[15] These authors contributed equally to this work.
[16] These authors jointly directed this work.

.

Nature Genetics: doi:10.1038/ng.3656

Extended Description of the Imputation Method with State Space Reduction

Here, we describe the state space reduction that uses the similarity between haplotypes in small genomic segments to reduce computational complexity. We recommend first reading a description of the original minimac algorithm[1]. Consider a reference panel with H haplotypes and a genomic segment bounded by markers P and Q. Let U ≤ H be the number of distinct haplotypes in the block. Label the original haplotypes as $X_1, X_2, \ldots, X_H$, and distinct unique haplotypes as $Y_1, Y_2, \ldots, Y_U$. For example, in **Figure 1**, the block B bounded by markers P=1 and Q=6 has U=3 distinct haplotypes.

Forward Equations

Let $L_k(.)$ and $\mathcal{L}_k(.)$ denote the left probabilities[2] for the original states and reduced states at marker k[2] ($P \leq k \leq Q$). Assuming we know $L_P(X_1), \ldots, L_P(X_H)$, equation (1) allows us to obtain $\mathcal{L}_P(Y_i)$ for each distinct haplotype.

$$\mathcal{L}_P(Y_i) = \sum_{\substack{j=1,\ldots,H \\ \text{and } X_j = Y_i}} L_P(X_j) \qquad (1)$$

In this reduced state space, we modify Baum-Welch's forward equations[3] to obtain $\mathcal{L}_k(.)$ recursively for $k = P+1, P+2, \ldots, Q$:

$$\mathcal{L}_{k+1}(Y_i) = \left[[1-\theta_k]\mathcal{L}_k(Y_i) + \frac{N_i\theta_k}{H} \sum_{j=1,\ldots,U} \mathcal{L}_K(Y_j) \right] \times P(S_{k+1}|Y_i) \qquad (2)$$

In (2), θ_k denotes the template switch probability between markers k and k+1 (analogous to a recombination fraction), S_{k+1} the genotype in the study sample, $P(S_{k+1}|Y_i)$ the genotype emission probabilities, and N_i the number of haplotypes matching Y_i in the original state space ($\sum_{i=1}^{U} N_i = H$). Once we obtain $\mathcal{L}_Q(.)$ for all the reduced states, we use them to calculate $L_Q(X_j)$ at the final block boundary, enabling us to transition between blocks. To accomplish this, we split probability $\mathcal{L}_Q(.)$ into two parts, $\mathcal{L}_Q^{NR}(.)$ and $\mathcal{L}_Q^{R}(.)$, where $\mathcal{L}_Q^{NR}(.)$ denotes the left probability at marker Q when no template switches occur between P and Q and $\mathcal{L}_{\mathrm{Q}}^{\mathrm{R}}(.)$ the probabilities when at least one switch occurs. This leads to equation (3) (where 'i' is such that $\mathrm{Y_i} = \mathrm{X_j}$):

$$L_Q(X_j) = \mathcal{L}_Q^R(Y_i)\times\left[\frac{1}{N_i}\right] + \mathcal{L}_Q^{NR}(Y_i)\left[\frac{L_P(X_j)}{\mathcal{L}_P(Y_i)}\right] \tag{3}$$

$\mathcal{L}_k^{NR}(.)$ and $\mathcal{L}_k^{R}(.)$ are defined as follows (for each k):

$$\mathcal{L}_k^{NR}(Y_i) = \mathcal{L}_P(Y_i)\prod_{i=P}^{k-1}[(1-\theta_i)P(S_{i+1}|Y_i)] \tag{4}$$

$$\mathcal{L}_k^{R}(Y_i) \ = \mathcal{L}_k(Y_i) - \mathcal{L}_k^{NR}(Y_i) \tag{5}$$

Backward Equations

Similar equations can be derived for the right probabilities $\mathrm{R_k}(.)$ and $\mathcal{R}_{\mathrm{k}}(.)$: equation (6) transforms the right probabilities ($\mathrm{R_Q} \rightarrow \mathcal{R}_{\mathrm{Q}}$), (7) gives the modified formulation for the Baum-Welch backward equations ($\mathcal{R}_{\mathrm{Q}} \rightarrow \mathcal{R}_{\mathrm{P}}$), and (8) transforms back the right probabilities ($\mathcal{R}_{\mathrm{P}} \rightarrow \mathrm{R_P}$).

$$\mathcal{R}_{\mathrm{Q}}(\mathrm{Y_i}) = \sum_{\substack{\mathrm{j=1,\dots,H} \\ \mathrm{and\ X_j=Y_i}}} \mathrm{R_Q}(\mathrm{X_j}) \tag{6}$$

$$\mathcal{R}_{\mathrm{k}}(\mathrm{Y_i}) = \frac{\mathrm{N_i}\theta_{\mathrm{k}}}{\mathrm{H}}\left[\sum_{\mathrm{j=1}}^{\mathrm{U}} \mathcal{R}_{\mathrm{k+1}}(\mathrm{Y_j})\mathrm{P}(\mathrm{S_{k+1}}|\mathrm{Y_j})\right] + \ [(1-\theta_{\mathrm{k}})\mathcal{R}_{\mathrm{k+1}}(\mathrm{Y_i})\mathrm{P}(\mathrm{S_{k+1}}|\mathrm{Y_i})] \tag{7}$$

$$R_P(X_j) = \mathcal{R}_P^R(Y_i)\times\left[\frac{1}{N_i}\right] + \mathcal{R}_P^{NR}(Y_i)\left[\frac{R_Q(X_j)}{\mathcal{R}_Q(Y_i)}\right] \tag{8}$$

$\mathcal{R}_k^{NR}(.)$ and $\mathcal{R}_k^{R}(.)$ are defined as follows (for each k):

$$\mathcal{R}_k^{NR}(Y_i) = \mathcal{R}_Q(Y_i)\prod_{i=k}^{Q-1}[(1-\theta_i)P(S_{i+1}|Y_i)] \tag{9}$$

$$\mathcal{R}_k^{R}(Y_i) \ = \mathcal{R}_k(Y_i) - \mathcal{R}_k^{NR}(Y_i) \tag{10}$$

Final Imputation Formula

Once we have the left and right probabilities for all the reduced states, the posterior probabilities for a template including any allele of interest at marker k can be calculated within the reduced state space as:

$$\mathrm{P}(\mathrm{Y_i}) = \left[\sum_{\substack{\mathrm{j=1,\dots,H} \\ \mathrm{and\ X_j=Y_i}}} \mathrm{L_P}(\mathrm{X_j})\mathrm{R_Q}(\mathrm{X_j})\right] \times \left[\frac{\mathcal{L}_\mathrm{K}^\mathrm{NR}(\mathrm{Y_i})}{\mathcal{L}_\mathrm{P}(\mathrm{Y_i})} \times \frac{\mathcal{R}_\mathrm{K}^\mathrm{NR}(\mathrm{Y_i})}{\mathcal{R}_\mathrm{Q}(\mathrm{Y_i})}\right] + \frac{1}{\mathrm{N_i}}\left[\mathcal{L}_\mathrm{K}(\mathrm{Y_i})\mathcal{R}_\mathrm{K}(\mathrm{Y_i}) - \mathcal{L}_\mathrm{K}^\mathrm{NR}(\mathrm{Y_i})\mathcal{R}_\mathrm{K}^\mathrm{NR}(\mathrm{Y_i})\right] \qquad (11)$$

Derivations of Formulations

Here, we prove that the formulations for the reduced state space HMM are mathematically equivalent to the original HMM. First we prove equation (3) ($\mathcal{L}_\mathrm{Q} \rightarrow \mathrm{L_Q}$), which states that the left probabilities of the original states can be extracted from the left probabilities of the reduced states (the proof is similar for the right probabilities).

Claim: For any K such that (P ≤ K ≤ Q) and $\mathrm{X_j}$ such that $\mathrm{X_j = Y_i}$

$$\mathrm{L_K}(\mathrm{X_j}) = \mathcal{L}_\mathrm{K}^\mathrm{R}(\mathrm{Y_i}) \times \left[\frac{1}{\mathrm{N_i}}\right] + \mathcal{L}_\mathrm{K}^\mathrm{NR}(\mathrm{Y_i}) \left[\frac{\mathrm{L_P}(\mathrm{X_j})}{\mathcal{L}_\mathrm{P}(\mathrm{Y_i})}\right] \qquad (12)$$

Proof:

We use mathematical induction to prove this claim. Proving it for K=P+1 is trivial (follows easily from the general proof given below). To prove it for general K>P), we show that the expression of $\mathrm{L_K}(\mathrm{X_j})$ from equation (12) satisfies the actual recursion for the forward equations in the original HMM[1]:

$$\mathrm{L_K}(\mathrm{X_j}) = \left[[1-\theta_{\mathrm{K}-1}]\mathrm{L_{K-1}}(\mathrm{X_j}) + \frac{\theta_{\mathrm{k}-1}}{\mathrm{H}}\sum_{\mathrm{i=1}}^{\mathrm{H}} \mathrm{L_{K-1}}(\mathrm{X_i})\right] \times \mathrm{P}(\mathrm{S_K}|\mathrm{X_j}) \qquad (13)$$

We first note that equation (4) can be re-written as follows:

$$\mathcal{L}_K^{NR}(Y_i) = \mathcal{L}_{K-1}^{NR}(Y_i)(1-\theta_{K-1})P(S_K|Y_i)$$

Accordingly, equation (5) becomes on substituting expression for $\mathcal{L}_k(Y_i)$ from equation (2):

$$\mathcal{L}_K^R(Y_i) \;=\left[[1-\theta_{K-1}]\mathcal{L}_{K-1}^R(Y_i)+\frac{N_i\theta_{K-1}}{H}\sum_{j=1}^{U}\mathcal{L}_{K-1}(Y_j)\right]\times P(S_K|Y_i)$$

Substituting the values of $\mathcal{L}_K^R(Y_i)$ and $\mathcal{L}_K^{NR}(Y_i)$ from the above equations in the RHS of equation (12) we get

$$\mathrm{RHS}=\mathcal{L}_{\mathrm{K}}^{\mathrm{R}}(\mathrm{Y_i})\times\left[\frac{1}{\mathrm{N_i}}\right]+\mathcal{L}_{\mathrm{K}}^{\mathrm{NR}}(\mathrm{Y_i})\left[\frac{\mathrm{L_P}(\mathrm{X_j})}{\mathcal{L}_{\mathrm{P}}(\mathrm{Y_i})}\right]$$

$$=\left[\frac{(1-\theta_{\mathrm{K}-1})\mathcal{L}_{\mathrm{K}-1}^{\mathrm{R}}(\mathrm{Y_i})}{\mathrm{N_i}}+\frac{\theta_{\mathrm{K}-1}\sum_{\mathrm{j}=1}^{\mathrm{U}}\mathcal{L}_{\mathrm{K}-1}(\mathrm{Y_j})}{\mathrm{H}}+\frac{\mathcal{L}_{\mathrm{K}-1}^{\mathrm{NR}}(\mathrm{Y_i})(1-\theta_{\mathrm{K}-1})\mathrm{L_P}(\mathrm{X_j})}{\mathcal{L}_{\mathrm{P}}(\mathrm{Y_i})}\right]\times\mathrm{P}(\mathrm{S_K}|\mathrm{Y_i})$$

$$=\left[(1-\theta_{\mathrm{K}-1})\left[\mathcal{L}_{\mathrm{K}-1}^{\mathrm{R}}(\mathrm{Y_i})\left[\frac{1}{\mathrm{N_i}}\right]\;+\mathcal{L}_{\mathrm{K}-1}^{\mathrm{NR}}(\mathrm{Y_i})\left[\frac{\mathrm{L_P}(\mathrm{X_j})}{\mathcal{L}_{\mathrm{P}}(\mathrm{Y_i})}\right]\right]+\frac{\theta_{\mathrm{K}-1}\sum_{\mathrm{j}=1}^{\mathrm{U}}\mathcal{L}_{\mathrm{K}-1}(\mathrm{Y_j})}{\mathrm{H}}+\right]\times\mathrm{P}(\mathrm{S_K}|\mathrm{Y_i})$$

$$=\left[(1-\theta_{\mathrm{K}-1})[\mathrm{L}_{\mathrm{K}-1}(\mathrm{X_j})]+\frac{\theta_{\mathrm{K}-1}\sum_{\mathrm{i}=1}^{\mathrm{H}}\mathrm{L}_{\mathrm{K}-1}(\mathrm{X_i})}{\mathrm{H}}+\right]\times\mathrm{P}(\mathrm{S_K}|\mathrm{X_j})$$

$$=\mathrm{L_K}(\mathrm{X_j})=\mathrm{LHS}$$

The last step follows from the induction hypothesis (i.e. $\mathrm{L}_{\mathrm{K}-1}(\mathrm{X_j})=\mathcal{L}_{\mathrm{K}-1}^{\mathrm{R}}(\mathrm{Y_i})\left[\frac{1}{\mathrm{N_i}}\right]+\mathcal{L}_{\mathrm{K}-1}^{\mathrm{NR}}(\mathrm{Y_i})\left[\frac{\mathrm{L_P}(\mathrm{X_j})}{\mathcal{L}_{\mathrm{P}}(\mathrm{Y_i})}\right]$) and from the identity $\sum_{\mathrm{i}=1}^{\mathrm{H}}\mathrm{L}_{\mathrm{K}-1}(\mathrm{X_i})=\sum_{\mathrm{j}=1}^{\mathrm{U}}\mathcal{L}_{\mathrm{K}-1}(\mathrm{Y_j})$ which follows from equation (1).

Next, we prove equation (11) which claims that the posterior probabilities obtained from the reduced states would be numerically same as those obtained from the original state space, proving that both HMMs are mathematically equivalent.

Claim: The posterior probability of each reduced state $\mathrm{Y_i}$ is given as:

$$\mathrm{P}(\mathrm{Y_i}) = \left[\sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \mathrm{L_P}(\mathrm{X_j})\mathrm{R_Q}(\mathrm{X_j})\right] \times \left[\frac{\mathcal{L}_K^{NR}(Y_i)}{\mathcal{L}_P(Y_i)} \times \frac{\mathcal{R}_K^{NR}(Y_i)}{\mathcal{R}_Q(Y_i)}\right] + \frac{1}{N_i}\left[\mathcal{L}_K(Y_i)\mathcal{R}_K(Y_i) - \mathcal{L}_K^{NR}(Y_i)\mathcal{R}_K^{NR}(Y_i)\right] \tag{14}$$

Proof:

To prove this, we start from the LHS of the above equation:

$$\mathrm{P}(\mathrm{Y_i}) = \sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \mathrm{P}(\mathrm{X_j})$$

$$= \sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \mathrm{L_K}(\mathrm{X_j})\mathrm{R_K}(\mathrm{X_j})$$

$$= \sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \left(\mathcal{L}_K^{R}(Y_i)\times\left[\frac{1}{N_i}\right] + \mathcal{L}_K^{NR}(Y_i)\left[\frac{L_P(X_j)}{\mathcal{L}_P(Y_i)}\right]\right)\times\left(\mathcal{R}_K^{R}(Y_i)\times\left[\frac{1}{N_i}\right] + \mathcal{R}_K^{NR}(Y_i)\left[\frac{R_Q(X_j)}{\mathcal{R}_Q(Y_i)}\right]\right)$$

$$= \left[\sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \mathrm{L_P}(\mathrm{X_j})\mathrm{R_Q}(\mathrm{X_j})\right] \times \left[\frac{\mathcal{L}_K^{NR}(Y_i)}{\mathcal{L}_P(Y_i)} \times \frac{\mathcal{R}_K^{NR}(Y_i)}{\mathcal{R}_Q(Y_i)}\right] + \sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \frac{\mathcal{L}_K^{R}(Y_i)\mathcal{R}_K^{R}(Y_i)}{N_i^2}$$

$$+ \frac{\mathcal{L}_K^{NR}(Y_i)\mathcal{R}_K^{R}(Y_i)}{N_i} \sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \frac{L_P(X_j)}{\mathcal{L}_P(Y_i)} + \frac{\mathcal{L}_K^{R}(Y_i)\mathcal{R}_K^{NR}(Y_i)}{N_i} \sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \frac{R_Q(X_j)}{\mathcal{R}_Q(Y_i)}$$

$$= \left[\sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \mathrm{L_P}(\mathrm{X_j})\mathrm{R_Q}(\mathrm{X_j})\right] \times \left[\frac{\mathcal{L}_K^{NR}(Y_i)}{\mathcal{L}_P(Y_i)} \times \frac{\mathcal{R}_K^{NR}(Y_i)}{\mathcal{R}_Q(Y_i)}\right]$$

$$+ \frac{1}{N_i}\left[\mathcal{L}_K^{R}(Y_i)\mathcal{R}_K^{R}(Y_i) + \mathcal{L}_K^{NR}(Y_i)\mathcal{R}_K^{R}(Y_i) + \mathcal{L}_K^{R}(Y_i)\mathcal{R}_K^{NR}(Y_i)\right]$$

$$= \left[\sum_{\substack{j=1,\dots,H\\ \text{and } X_j=Y_i}} \mathrm{L_P}(\mathrm{X_j})\mathrm{R_Q}(\mathrm{X_j})\right] \times \left[\frac{\mathcal{L}_K^{NR}(Y_i)}{\mathcal{L}_P(Y_i)} \times \frac{\mathcal{R}_K^{NR}(Y_i)}{\mathcal{R}_Q(Y_i)}\right] + \frac{1}{N_i}\left[\mathcal{L}_K(Y_i)\mathcal{R}_K(Y_i) - \mathcal{L}_K^{NR}(Y_i)\mathcal{R}_K^{NR}(Y_i)\right]$$

$$= \mathrm{LHS}$$

m3vcf Format Description

The m3vcf format is based on the VCF format (https://samtools.github.io/hts-specs/VCFv4.2.pdf) and applies the idea of state space reduction to store large reference panels using less disk space. m3vcf files save each genomic segment in series where each segment has the list of bi- and multi-allelic variants in order along with the unique haplotypes at these variants and a single line at the beginning of the block that describes which individual maps to which unique haplotype.

Example

```
##fileformat=M3VCF
##version=1.1
##compression=block
##n_blocks=2
##n_haps=12
##n_markers=8
##<Note=This is NOT a VCF File and cannot be read by vcftools>
#CHROM POS   ID            REF   ALT QUAL FILTER INFO                 FORMAT A1 A2 B1 B2 C1 C2 D1 D2 E1 E2 F1 F2
6      73924  <BLOCK:0-5>  .     .   .    .      B1;VARIANTS=6;REPS=4 .      0  1  3  0  0  0  1  0  3  1  0  3
6      73924  chr6:73924:D AAGAG A   .    .                            0000
6      89919  chr6:89919   T     G   .    .                            0100
6      89921  chr6:89921   C     T   .    .                            0000
6      89932  chr6:89932   A     G   .    .                            0000
6      89949  chr6:89949   G     A,T .    .                            0122
6      100116 chr6:100116  C     A   .    .                            0001
6      100116 <BLOCK:5-7>  .     .   .    .      B2;VARIANTS=3;REPS=2 .      0  1  0  0  0  0  1  0  1  1  0  1
6      100116 chr6:100116  T     A   .    .                            00
6      132285 chr6:132285  T     A,G .    .                            02
6      148689 chr6:148689  TAA   T   .    .                            01
```

Description

File meta-information starts with “##” and includes file format (fileformat), version (version), compression method (compression), number of genomic segments (n_blocks), number of haplotypes (n_haps), and number of markers (n_markers).

The header line starts with “#” and follows the VCF format definition.

The data lines define each genomic segment (denoted by <BLOCK:*-*>) followed by the markers contained in this genomic segment (denoted by their original marker IDs). In the example above, a reference panel of 6 samples (12 haplotypes) and 8 markers was reduced to two genomic segments (<BLOCK:0-5> and <BLOCK:5-7>). The first block from marker 0 to 5 (6 variants) and the second one from marker 5 to 7 (3 variants). Note that two consecutive blocks must overlap at one common marker. The INFO column stores the block number (Bx), the number of markers in a segment (VARIANTS), and the number of unique haplotypes in that segment (REPS). The following columns represent the unique label assigned to each sample in that block. The numbers for each sample represent the unique haplotype which resembles that genomic segment. In the data lines followed by the block definition, the details of the variants are stored along with the unique haplotypes in the FORMAT column. For example, for the <BLOCK:0-5>, we have 4 unique haplotypes (given by the variable REPS) which are the four sub-columns (of 0's and 1's) in the FORMAT column.

Source code to generate m3vcf files is included in minimac3 (http://genome.sph.umich.edu/wiki/Minimac3). Utilities to manipulate m3vcf files can be found here: http://genome.sph.umich.edu/wiki/M3vcftools.

Tool Command Lines

Table 1 command line parameters used for each imputation tool.

minimac3 (Version 1.0.14):

```
Minimac3            --refHaps $REF.m3vcf.gz
                    --haps $GWAS.vcf.gz
                    --chr 20
                    --start $START
                    --end $END
                    --doseOutput
                    --vcfOutput
                    --window 1000000
                    --prefix $OUTPUT
```

minimac2 (RELEASE STAMP 2014-05-12):

```
minimac-fst   --refHaps $REF.vcf.gz
                    --vcfReference
                    --haps $GWAS.hap
                    --snps $GWAS.snps
                    --em
                    --chr 20
                    --vcfstart $START
                    --vcfend $END
                    --vcfwindow 1000000
                    --prefix $OUTPUT
                    --rec $RECOM.rec
                    --erate $ERATE.erate
                    --rounds 0
```

IMPUTE2 (Version 2.3.1):

```
impute2             - known_haps_g $GWAS.hap.gz
```

```
                -h $REF.hap.gz
                -l $REF.legend.gz
                -m genetic_map_chr20_combined_b37.txt
                -int $START $END
                -Ne 2000
                -k_hap 500000
                -buffer 1000
                -o $OUTPUT
```

Beagle4.1 (RELEASE STAMP 22Feb 2016):

```
java -jar       gt=$GWAS.vcf.gz
                map= plink.chr20.GRCh37.map
                ref=$REF.bref
                window=23000
                overlap=4000
                nthreads=1
                niterations=0
                gprobs=true
```

References

1. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).
2. 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**, 816-34 (2010).
3. Baum, L.E., Petrie, T., Soules, G. & Weiss, N. A maximization technique occurring in statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics* **41**, 164-171 (1970).