Supplemental Figure 1

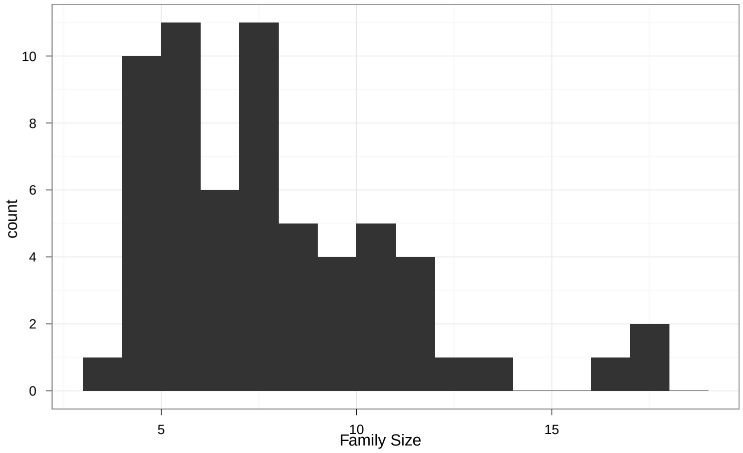

Figure 1: Distribution of the number of individuals in each family.

Supplemental Figure 2

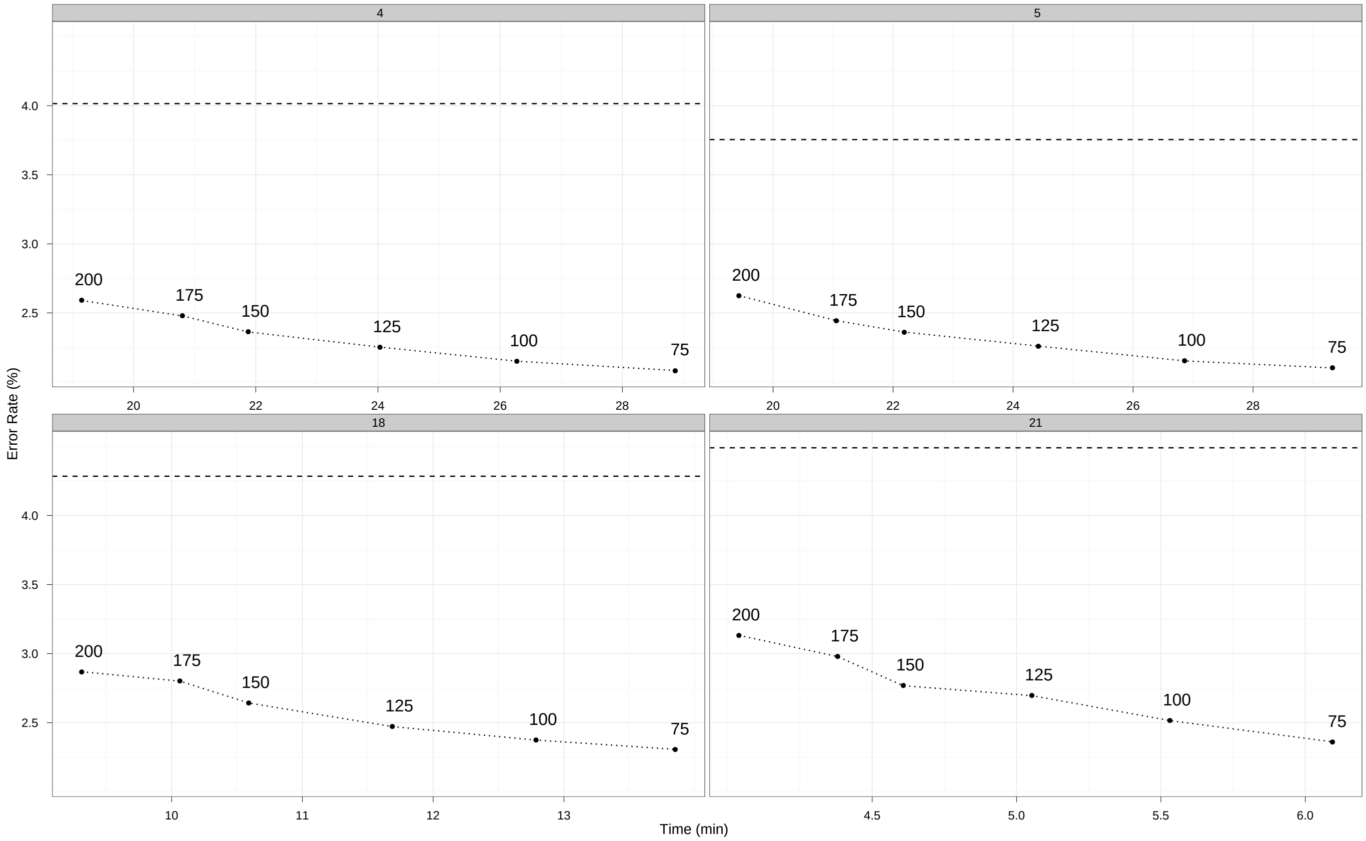

Figure 2: Accuracy versus time tradeoff for the Nesterov Algorithm on chromosomes 4, 5, 18 and 21 from the Chinese Han group in HapMap3. The numbers indicate the sub-window size w. The dashed line marks the error rate for MaCH on the same data set.

Supplemental Figure 3

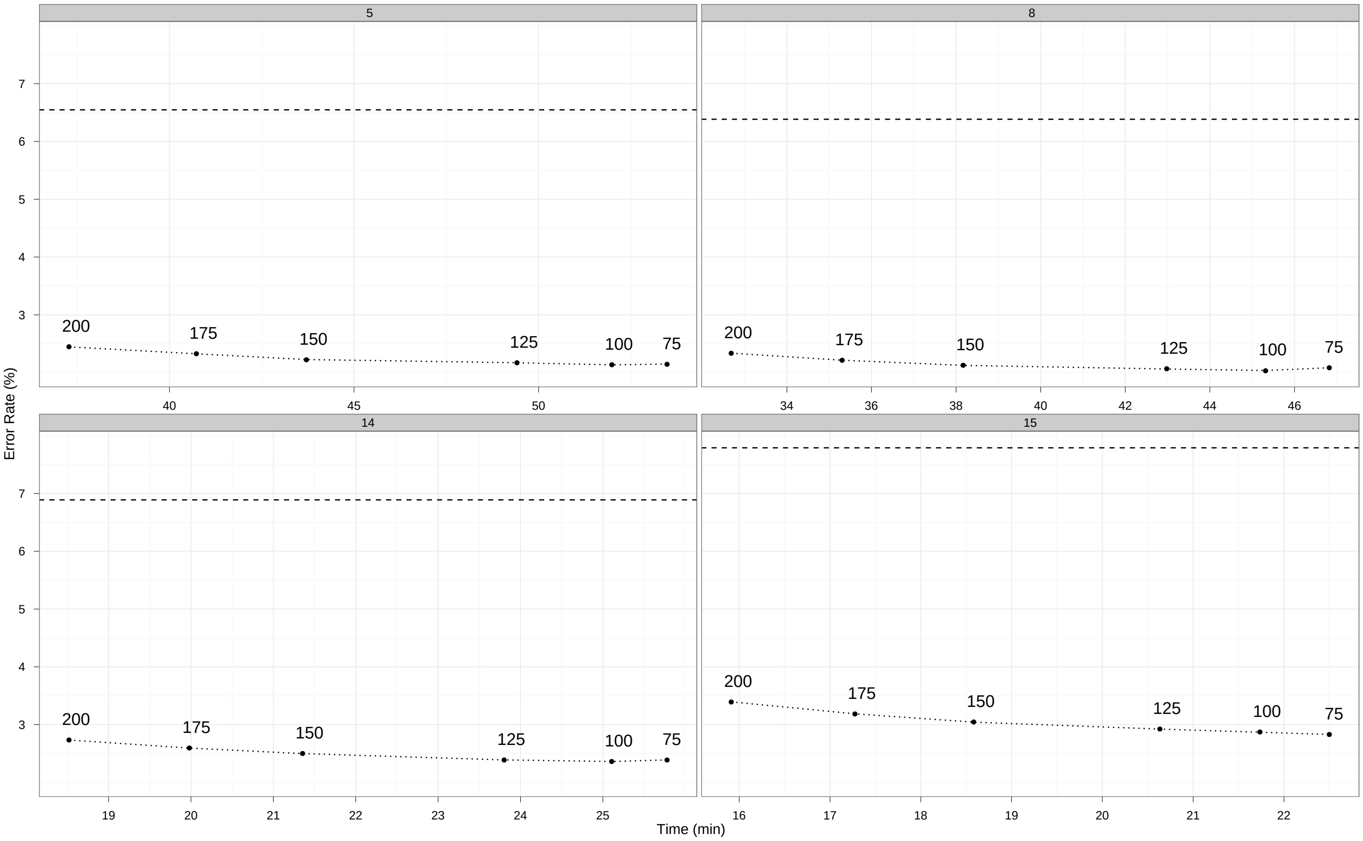

Figure 3: Accuracy versus time tradeoff for the Nesterov Algorithm on chromosomes 5, 8, 14 and 15 from the Yoruba group in HapMap3. The numbers indicate the sub-window size w. The dashed line marks the error rate for MaCH on the same data set.

Supplemental Figure 4

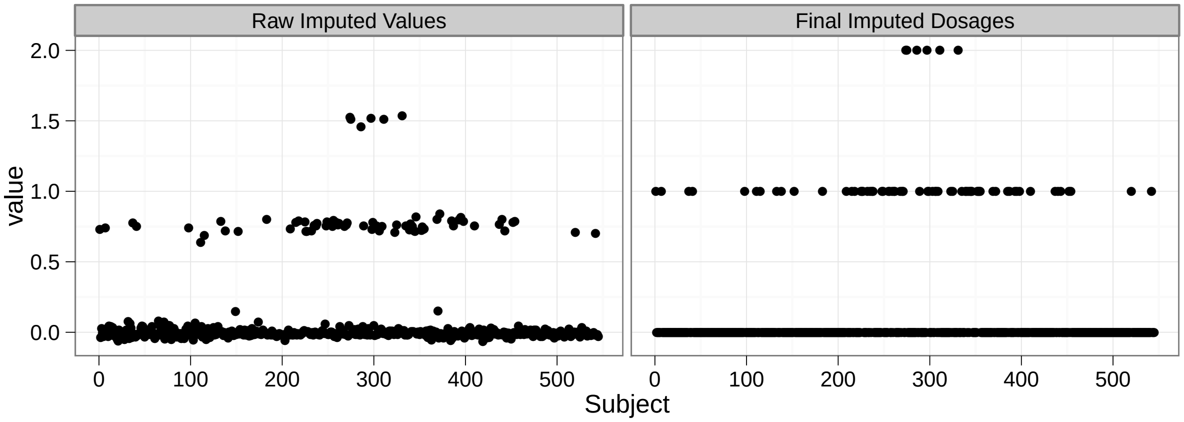

Figure 4: Raw Mendel-Impute output and final imputed dosages after EM clustering for an untyped SNP from the simulated Illumina experiment.The dosage is the posterior mean of the reference allele count for the given SNP. Note that the nuclear norm regularization outputs raw imputed values that are biased towards zero. On the other hand, EM clustering is able to successfully restore very reasonable dosages on the correct scale between zero and two.

Supplemental Note 1

In this section, we describe a competing MM algorithm to Nesterov's method and present numerical results showing the superiority of the latter.

MM algorithm

Recall that we seek to solve the following optimization problem

$$\min \; f(\mathbf{Z}) + \lambda \|\mathbf{Z}\|_*.$$

Given the current iterate $\mathbf{Z}^k$, the MM algorithm capitalizes on the function

$$Q(\mathbf{Z} \mid \mathbf{Z}^k) \;\; = \;\; \frac{1}{2}\sum_{i=1}^{p} \|P_{\Omega_i}(\mathbf{X}^i) + P_{\Omega_i}^{\perp}(\mathbf{Z}^k) - \mathbf{Z}\|_{\mathrm{F}}^2,$$

majorizing the loss $f(\mathbf{Z})$. Here

$$P_{\Omega}^{\perp}(\mathbf{Y})_{ij} \;\; = \;\; \begin{cases} y_{ij} & \text{if } (i,j) \notin \Omega, \\ 0 & \text{otherwise} \end{cases}$$

denotes the projection operator orthogonal to $P_{\Omega}(\mathbf{Y})$. Majorization is understood to mean

$$Q(\mathbf{Z}^k \mid \mathbf{Z}^k) \;\; = \;\; f(\mathbf{Z}^k) \quad \text{and} \quad Q(\mathbf{Z} \mid \mathbf{Z}^k) \;\; \geq \;\; f(\mathbf{Z}).$$

To prove these tangency conditions, one simply notes that $Q(\mathbf{Z} \mid \mathbf{Z}^k)$ adds back the missing square terms that distinguish $\|P_{\Omega_i}(\mathbf{X}^i) - P_{\Omega_i}^{\perp}(\mathbf{Z})\|_{\mathrm{F}}^2$ from $\|\mathbf{X}^i - \mathbf{Z}\|_{\mathrm{F}}^2$ and forces them to equal 0 when $\mathbf{Z} = \mathbf{Z}^k$. If we complete the square of $\|P_{\Omega_i}(\mathbf{X}^i) - P_{\Omega_i}^{\perp}(\mathbf{Z})\|_{\mathrm{F}}^2$, then we can rewrite $Q(\mathbf{Z} \mid \mathbf{Z}^k)$ as

$$Q(\mathbf{Z} \mid \mathbf{Z}^k) \;\; = \;\; \frac{1}{2}\sum_{i=1}^{p} \|P_{\Omega_i}(\mathbf{X}^i) + P_{\Omega_i}^{\perp}(\mathbf{Z}^k) - \mathbf{M}^k\|_{\mathrm{F}}^2 + \frac{p}{2}\|\mathbf{M}^k - \mathbf{Z}\|_{\mathrm{F}}^2,$$

where

$$\mathbf{M}^k \;\; = \;\; \frac{1}{p}\sum_{i=1}^{p} P_{\Omega_i}(\mathbf{X}^i) + \frac{1}{p}\sum_{i} P_{\Omega_i}^{\perp}(\mathbf{Z}^k).$$

For computational efficiency, the first term defining $\mathbf{M}^k$ should be pre-computed and stored. The second term is a Hadamard product $\mathbf{W} * \mathbf{Z}^k$, where entry w_{jk} of $\mathbf{W}$ reduces to the proportion $1 - p^{-1}\sum_{i=1}^{p} 1_{\{(j,k)\in\Omega_i\}}$ of platforms lacking typing on the person-SNP pair (j,k).

The MM algorithm minimizes the regularized surrogate function

$$Q(\mathbf{Z} \mid \mathbf{Z}^k) + \lambda\|\mathbf{Z}\|_* \;\; = \;\; \frac{p}{2}\|\mathbf{M}^k - \mathbf{Z}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{Z}\|_* + c^k$$

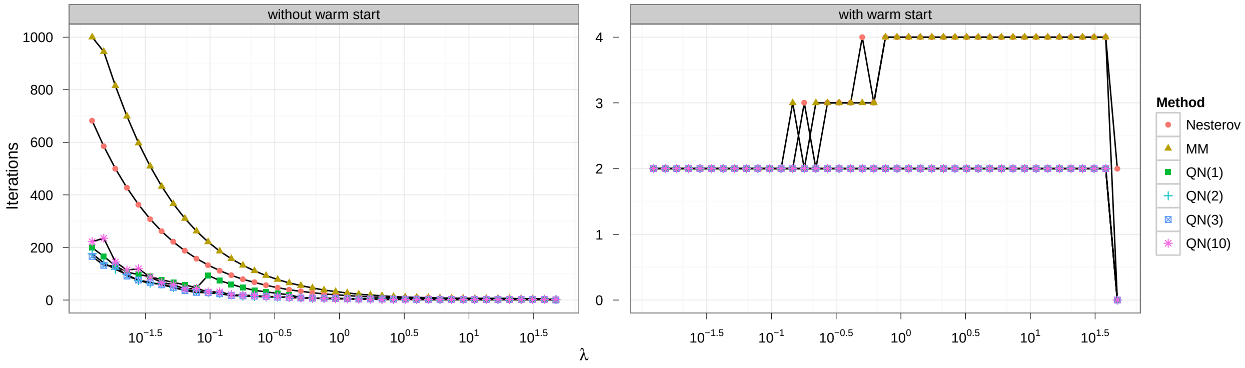

Figure 5: Comparison of iteration counts for matrix completion methods over a grid of regularization parameters on the Daly data set.

with respect to $\mathbf{Z}$. Here the constant $c^k = \frac{1}{2}\sum_{i=1}^{p} \|P_{\Omega_i}(\mathbf{X}^i) + P_{\Omega_i}^{\perp}(\mathbf{Z}^k) - \mathbf{M}^k\|_{\mathrm{F}}^2$ is irrelevant. Minimization gives the next iterate $\mathbf{Z}^{k+1}$ and drives the penalized loss (1) downhill. Minimization of the surrogate function relies on two facts. First, the penalty involves only the singular values of the matrix $\mathbf{M}^k$. Second, the surrogate loss $\frac{p}{2}\|\mathbf{M}^k - \mathbf{Z}\|_{\mathrm{F}}^2$ is minimized by aligning the eigenstructure of the matrix $\mathbf{Z}$ with the eigenstructure of the matrix $\mathbf{M}^k$, regardless of the singular values of $\mathbf{Z}$. Thus, if $\mathbf{M}^k$ admits the singular value decomposition $\mathbf{M}^k = \mathbf{U}\mathrm{diag}(\mathbf{m}^k)\mathbf{V}^T$, then the optimal $\mathbf{Z}^{k+1}$ shares its singular vectors with $\mathbf{M}^k$ (Lange, 2012). The singular values $z_i^{k+1} = (m_i^k - \lambda/p)_+$ of $\mathbf{Z}$ are shrunken versions of the singular values of $\mathbf{M}^k$. The pressure exerted by the nuclear norm penalty forces this shrinkage. The resulting procedure is summarized in Algorithm 1. Standard theory for the MM algorithm (Lange, 2010) shows that $\mathbf{Z}^k$ monotonically converges to a global minimum of the objective function (1).

```
Pre-compute $\bar{\mathbf{X}} = p^{-1}\sum_i P_{\Omega_i}(\mathbf{X}^i)$ and $\mathbf{W} = (w_{jk}) = (1 - p^{-1}\sum_i 1_{\{(j,k)\in\Omega_i\}})$ ;
Initialize $\mathbf{Z}^0$ ;
repeat
    $\mathbf{M}^k \leftarrow \bar{\mathbf{X}} + \mathbf{W} * \mathbf{Z}^k$ ;
    Compute SVD $\mathbf{M}^k = \mathbf{U}\mathrm{diag}(\mathbf{m}^k)\mathbf{V}^T$ ;
    $\mathbf{z}^{k+1} \leftarrow (\mathbf{m}^k - \lambda/p)_+$ ;
    $\mathbf{Z}^{k+1} \leftarrow \mathbf{U}\mathrm{diag}(\mathbf{z}^{k+1})\mathbf{V}^T$ ;
until objective value converges;
```

Algorithm 1: MM algorithm for minimizing the penalized loss (1).

Comparison of Nesterov's method and the MM algorithm with the Daly Data

The preceding discussion suggests that the Nesterov method is preferable because it enjoys faster convergence than the MM algorithm. However, the question is complicated for two reasons. First,

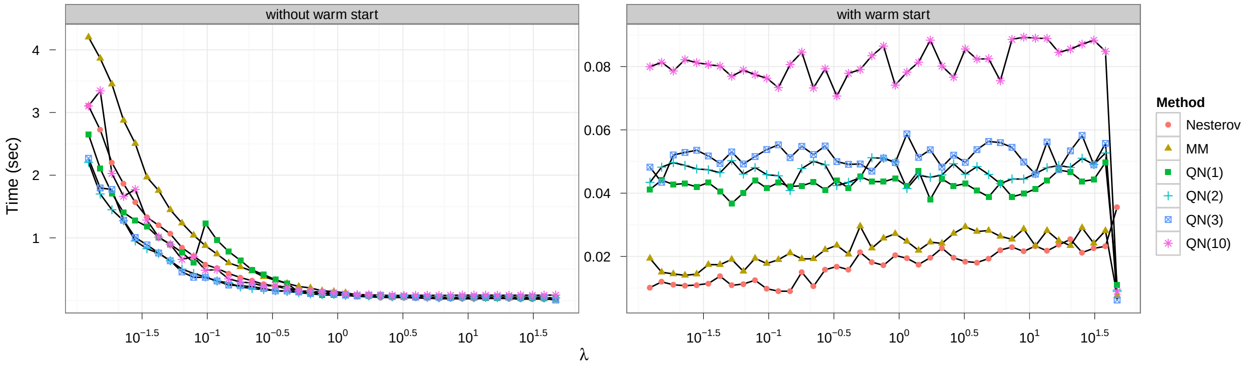

Figure 6: Comparison of run times (sec) for matrix completion methods over a grid of λ values on the Daly data.

since the MM algorithm is a fixed point algorithm, the quasi-Newton (QN) acceleration derived by (Zhou et al., 2011) could boost its performance significantly and should be explored. Second, any application of penalized estimation requires tuning a penalty constant such as λ. Therefore, algorithmic efficiency should be evaluated as the time to solve the optimization problem (1) over a grid of decreasing λ values. For large values of λ, the solution coincides with $\mathbf{Z}_\lambda = \mathbf{0}_{m\times n}$. Near the maximum singular value $\lambda_{\max}$ of $\bar{\mathbf{X}}$, the solution starts to diverge from this trivial value. Hence, the grid starts at $\lambda_{\max}$.

We evaluated the numerical speed and reliability of both methods on a toy data set (Daly et al., 2001). These data contain 103 SNPs on chromosome 5q31 genotyped on 387 parent-offspring trios. We consider only the 129 children; 8% of their genotypes are missing in the reduced data set. We report two sets of experiments. In both we compare Nesterov's method, the MM algorithm, and the quasi-Newton accelerated MM algorithm over a fixed set of 41 λ values. The quasi-Newton acceleration scheme relies on a user specified number of secant conditions. More secant conditions tend to improve the rate of convergence at the expense of additional computational complexity. In this test we use 1, 2, 3, and 10 secant conditions. In the first set of experiments, we compute the solution at λ value starting from the matrix $\mathbf{0}_{m\times n}$. In the second set of experiments, we evaluate the effect of warm starts, in which the solution computed at a given λ value is taken as the starting value for the computation of the solution at the next λ value. Figure 5 compares the number of iterations taken by each method; Figure 6 compares the time taken by each method.

Without warm starts, the left panel of Figure 5 clearly distinguishes the various methods for small values of λ. The MM algorithm does the worst in terms of iteration counts, followed by Nesterov's Method. The quasi-Newton scheme does the best over the range of secant conditions tested. Differences in computation times are less pronounced than differences in iteration counts because computing the quasi-Newton updates imposes an additional overhead. However, as the right panels of Figures 5 and 6 make evident, the tables are turned when warm starts are employed. Both the unaccelerated MM and Nesterov's method now outperform the quasi-Newton scheme, with Nesterov's method performing the best. Moreover, we see that warm starts drastically cut the required computational effort. It should be clear from these comparisons that the Nesterov algorithm with warm start is optimal. Indeed, being able to rapidly compute solutions over a range of λ is the key to the speed gains enjoyed by our model-free imputation method.

Supplemental Note 2

By default fastPHASE uses cross-validation to choose either 10 or 20 haploytpe clusters. CHB in general displays more linkage disequilibrium than YRI. With the exception of chromosome 5, fastPHASE chose 10 clusters, while it chose 20 for chromosome 5. fastPHASE chose 20 clusters for all YRI chromosomes. Thus, the exceptionally long run times for fastPHASE coincide when the algorithm was prompted by the data to use larger clusters.

Supplemental Note 3

We present two toy problems to give some intuition on when Mendel-Impute is expected to perform both poorly and well. Suppose we have a reference panel of four haplotypes of 3 SNPs with equal frequency in the population panel: 101, 110, 011, 000. We constructed a study panel of 32 individuals by randomly sampling with replacement two of the four haplotypes for each individual. We masked the third SNP. Using a reference panel of all 16 unique possible genotypes, we applied Mendel-Impute to impute the untyped third SNP. We performed 100 replicates of this scenario. We repeated the same experiment using four haplotypes of 6 SNPs again with equal frequency in the population panel: 101100, 110010, 011001, 000111. For the former we used a fixed regularization parameter that resulted in a rank 3 approximation for the first scenario and a rank 4 approximation in the second for all replicates. Table 1 shows that Mendel-Impute struggled with imputing the first scenario but enjoyed great success in the second scenario.

We should not be surprised that Mendel-Impute struggled on the first problem and performed well on the second. In the first problem, the underlying haplotype variation is too rich for a matrix completion framework to capture. There are four underlying signals, suggesting a rank 4 approximation, but matrix completion is limited to finding at most a rank 3 approximation

Number of SNPs in haplotype	3	6
Min	0	18
Mean	5.79	31.72
Median	3	32
Max	17	32

Table 1: Summary statistics for 100 replicates on the number of correctly imputed SNPs for Mendel-Impute applied to 32 subjects with genotypes simulated from four haplotypes of length 3 and 6 SNPs. The third SNP was missing in both scenarios.

since the study panel matrix of genotypes is 48-by-3. The model is not flexible enough to capture the variation in the data. In the second problem, however, the matrix completion framework is limited to finding at most a rank 6 approximation since the study panel matrix is now 48-by-6. Indeed, the rank 4 approximation, that was selected, is able to capture enough of the systematic variation due to the underlying 4 haplotypes to correctly impute all the missing SNPs. The two toy problems demonstrate that in order for Mendel-Impute to work, we need the number of underlying haplotypes to be small relative to the number of SNPs in the haplotype block.

Supplemental Note 4

For the problem at hand $k = 3$. Making the following choices works well in practice. We initialize $\pi_1 = f^2, \pi_2 = 2f(1-f)$,and $\pi_3 = (1-f)^2$ where f is the reference allele frequency in say the reference panel. We set the parameters $\alpha_j = \alpha\pi_j + 1$ where α is very large, typically on the order of 10^{12} to apply a very strong prior. Let x_i denote the raw Mendel-Impute output for the i study subject for the SNP of interest. Then we first project x_i onto the interval $[0, 2]$. We then initialize $\mu_3 = \max_i x_i, \mu_1 = 0$,and $\mu_2 = (\mu_1 + \mu_2)/2$.

References

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S., 2001. High-resolution haplotype structure in the human genome. *Nat Genet*, **29**(2):229–232.

Lange, K., 2010. *Numerical Analysis for Statisticians.* Statistics and Computing. Springer, New York, second edition.

Lange, K., 2012. *Optimization.* Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Zhou, H., Alexander, D., and Lange, K., 2011. A quasi-newton acceleration for high-dimensional optimization algorithms. *Statistics and Computing*, **21**:261–273.