# Haplotype estimation for biobank scale datasets: Supplementary Note and Tables

Jared O'Connell, Kevin Sharp, Nick Shrine, Louise Wain, Ian Hall, Martin Tobin, Jean-Francois Zagury, Olivier Delaneau, Jonathan Marchini

May 4, 2016

## Contents

# 1 Algorithm for haplotype partioning

**Algorithm 1** Fast haplotype partitioning

```
$c$=0                                                      ▷ Unique cluster identifier
$d$=0                                                      ▷ Depth of tree
$Z = \{1 \dots N\}$                                        ▷ Set containing haplotype indices
PARTITION($Z$, $c$, $d$)

function PARTITION($Z$, $c$, $d$)
    $B =$ KMEANS($\{H_i : i \in Z\}$)                      ▷ Partition haplotypes into two clusters
    $Z^1 = \{Z_i : B_i = 1\}$                              ▷ Indices of cluster 1/2 stored in respective $Z$s
    $Z^2 = \{Z_i : B_i = 2\}$
    if $|Z^1| < M$ then                                    ▷ Terminate if cluster is small enough, otherwise recurse
        TERMINATE($Z^1$, $Z^2$, $c$)
    else
        PARTITION($Z^1$, $c$, $d$)
    end if
    if $|Z^2| < M$ then
        TERMINATE($Z^2$, $Z^1$, $c + 2^d$)
    else
        PARTITION($Z^2$, $c + 2^d$, $d$)
    end if
    $d++$
end function

function TERMINATE($Z^1$, $Z^2$, $c$)
    Calculate centroid $\bar{H}$ where $\bar{H}_l = \frac{1}{|Z^1|} \sum_{z \in Z^1} H_{z,l}$
    $Q = \frac{M - |Z^1|}{|Z^2|}$ quantile of $\{\texttt{euc}(H_{z_i}, \bar{H}) : z_i \in Z^2\}$
    $D[c] = Z^1 \cup \{z_i \in Z^2 : \texttt{euc}(H_{z_i}, \bar{H}) \leq Q\}$     ▷ Adds secondary cluster to dictionary
    for $z_i \in Z^1$ do
        $C[z_i] = c$                                       ▷ Assigns primary cluster
    end for
end function
```

## 2 A Hidden Markov Model for Sibling Pairs

Consider a pair of siblings with genotypes $G_\ell^a$ and $G_\ell^b \in \{0,1,2\}$ at genomic positions $\ell \in \{1,\ldots,L\}$ across a chromosome with $L$ markers. The value of $G_\ell^a$ indicates the number of minor alleles at the $\ell^{th}$ site for sibling a. We denote a pair of maternal alleles at position $\ell$ by $\left(m_\ell^1, m_\ell^2\right)$ and a pair of paternal alleles by $\left(f_\ell^1, f_\ell^2\right)$. The genotype, $G_\ell^a$, of each sibling arises from the combination of one maternal allele and one paternal allele. If sibling $a$ inherits $m_\ell^1$ and $f_\ell^2$ and sibling b inherits $m_\ell^1$ and $f_\ell^1$, we denote this combination as $\{\left(m_\ell^1, f_\ell^2\right)^a, \left(m_\ell^1, f_\ell^1\right)^b\}$. For two siblings there are 16 such combinations which we assume are equally probable.

The number of identical alleles shared by the siblings determines the IBD status at the $\ell^{th}$ site, $X_\ell \in \{0,1,2\}$. For example, the pair $\{\left(m_\ell^1, f_\ell^2\right)^a, \left(m_\ell^1, f_\ell^1\right)^b\}$ denotes the sharing of one identical allele, $m_\ell^1$ at site $\ell$, so $X_\ell = 1$. Typically, the IBD status is unobserved. However we can infer it from the observed IBS status.

The IBS status at the $\ell^{th}$ site, $Y_\ell \in \{0,1,2\}$, is determined by the number of pairs of alleles of the same type, major or minor (denoted by 0 and 1 respectively) shared by the two individuals. While $Y_\ell$ depends on $X_\ell$, they are not the same. For example, the combination $\{\left(m_\ell^1 = 0, f_\ell^2 = 1\right)^a, \left(m_\ell^1 = 0, f_\ell^1 = 1\right)^b\}$ has $X_\ell = 1$, but $Y_\ell = 2$ (a and b each have one minor allele and one major allele), while $\{\left(m_\ell^1 = 0, f_\ell^2 = 0\right)^a, \left(m_\ell^2 = 0, f_\ell^1 = 1\right)^b\}$ has $X_\ell = 0$, but $Y_\ell = 1$.

Our HMM is specified by an emission distribution, $P\left(Y_\ell | X_\ell\right)$, which captures the dependence of IBS status on IBD status, a prior distribution, $P\left(X_\ell\right)$, over IBD status and a set of transition probabilities, $P\left(X_\ell | X_{\ell-1}\right)$.

To compute the emission probabilities, $P\left(Y_\ell = y | X_\ell = x\right)$, we enumerate the mutually exclusive combinations of alleles consistent with $Y_\ell = y$ conditioned on $X_\ell = x$ and sum their respective probabilities. For example, if $Y_\ell = 0$ given $X_\ell = 0$ there are 8 possible combinations of alleles:

$$\begin{gathered}
\left\{\left(m_\ell^1 = 0, f_\ell^1 = 0\right)^a, \left(m_\ell^2 = 1, f_\ell^2 = 1\right)^b\right\}, \left\{\left(m_\ell^1 = 1, f_\ell^1 = 1\right)^a, \left(m_\ell^2 = 0, f_\ell^2 = 0\right)^b\right\} \\
\left\{\left(m_\ell^1 = 0, f_\ell^2 = 0\right)^a, \left(m_\ell^2 = 1, f_\ell^1 = 1\right)^b\right\}, \left\{\left(m_\ell^1 = 1, f_\ell^2 = 1\right)^a, \left(m_\ell^2 = 0, f_\ell^1 = 0\right)^b\right\} \\
\left\{\left(m_\ell^2 = 0, f_\ell^1 = 0\right)^a, \left(m_\ell^1 = 1, f_\ell^2 = 1\right)^b\right\}, \left\{\left(m_\ell^2 = 1, f_\ell^1 = 1\right)^a, \left(m_\ell^1 = 0, f_\ell^2 = 0\right)^b\right\} \\
\left\{\left(m_\ell^2 = 0, f_\ell^2 = 0\right)^a, \left(m_\ell^1 = 1, f_\ell^1 = 1\right)^b\right\}, \left\{\left(m_\ell^2 = 1, f_\ell^2 = 1\right)^a, \left(m_\ell^1 = 0, f_\ell^1 = 0\right)^b\right\}
\end{gathered}$$

To compute the probability of one of these possible combinations, we assume a uniform distribution over the four different combinations of transmitted alleles:

$$\begin{aligned}
P\left(\left\{\left(m_\ell^1, f_\ell^1\right)^a, \left(m_\ell^2, f_\ell^2\right)^b\right\} | X_\ell = 0\right) &= P\left(\left\{\left(m_\ell^2, f_\ell^1\right)^a, \left(m_\ell^1, f_\ell^2\right)^b\right\} | X_\ell = 0\right) = \ldots \\
&= P\left(\left\{\left(m_\ell^2, f_\ell^2\right)^a, \left(m_\ell^1, f_\ell^1\right)^b\right\} | X_\ell = 0\right) = \frac{1}{4}.
\end{aligned}$$

We also assume independence between the probabilities of each different allele being observed as major or minor and estimate the probability, $\alpha_\ell$, of observing a minor allele at position $\ell$ from its empirical frequency in the full $N = 49,458$ cohort. For example, if we denote the first of the eight combinations above by $C_1$ we would compute $P\left(C_1|X_\ell\right)$ as:

$$\begin{aligned} P\left(C_1|X_\ell\right) &= P\left(\left\{\left(m_\ell^1, f_\ell^1\right)^a, \left(m_\ell^2, f_\ell^2\right)^b\right\}|X_\ell = 0\right) P\left(\left\{\left(0,0\right), \left(1,1\right)\right\}|X_\ell = 0\right) \\ &= \frac{1}{4}\left(1-\alpha_\ell\right)^2 \alpha_\ell^2. \end{aligned}$$

By symmetry one can see that the probability of observing each of these eight, mutually exclusive combinations is the same. By summing, we obtain:

$$P\left(Y_\ell = 0|X_\ell = 0\right) = 2\left(1-\alpha_\ell\right)^2 \alpha_\ell^2$$

In cases where $X_\ell \neq 0$ the observed alleles are no longer all independent. For example, one combination consistent with the case $Y_\ell = 1$ conditioned on $X_\ell = 1$ would be:

$$C_2 = \left\{\left(m_\ell^1 = 0, f_\ell^1 = 0\right)^a, \left(m_\ell^1 = 0, f_\ell^2 = 1\right)^b\right\}.$$

There are eight different combinations of transmitted alleles consistent with $X_\ell = 1$, but, as $m_\ell^1$ is transmitted to both siblings, only three of the observed alleles are now independent. Hence, the probability of the combination $C_2$ would be given by:

$$\begin{aligned} P\left(C_2|X_\ell\right) &= P\left(\left\{\left(m_\ell^1, f_\ell^1\right)^a, \left(m_\ell^1, f_\ell^2\right)^b\right\}|X_\ell = 0\right) P\left(\left\{\left(0,0\right), \left(0,1\right)\right\}|X_\ell = 0\right) \\ &= \frac{1}{8}\left(1-\alpha_\ell\right)^2 \alpha_\ell. \end{aligned}$$

So far we have assumed no genotyping error. However, such errors can have an adverse effect on inference by introducing genotype configurations that are incompatible with the underlying true IBD status. Therefore, as an additional refinement, we incorporate a simple error model. We assume a small genotyping error rate $\epsilon = 0.005$, which is multiplied by the probability of an error introducing an incompatible genotype configuration. We further assume that such genotype errors result in a uniformly distributed random pair of genotypes. Thus, for example, instead of $P\left(Y_\ell = 0|X_\ell = 1\right) = 0$, we define $P\left(Y_\ell = 0|X_\ell = 1\right) = \frac{2}{9}\epsilon$.

The full set of emission distributions is summarised below:

$$
\begin{array}{l}
X_\ell = 0 \left\{ \begin{array}{ll}
P\left(Y_\ell = 0 | X_\ell = 0\right) & = 2\alpha_\ell^2 \left(1-\alpha_\ell\right)^2 , \\
P\left(Y_\ell = 1 | X_\ell = 0\right) & = 4\left(\alpha_\ell \left(1-\alpha_\ell\right)^3 + \alpha_\ell^3 \left(1-\alpha_\ell\right)\right) , \\
P\left(Y_\ell = 2 | X_\ell = 0\right) & = \alpha^4 + 4\alpha_\ell^2 \left(1-\alpha_\ell\right)^2 + \left(1-\alpha_\ell\right)^4 ,
\end{array} \right. \\
X_\ell = 1 \left\{ \begin{array}{ll}
P\left(Y_\ell = 0 | X_\ell = 1\right) & = \frac{2}{9}\epsilon, \\
P\left(Y_\ell = 1 | X_\ell = 1\right) & = 2\left(\alpha_\ell \left(1-\alpha_\ell\right)^2 + \alpha_\ell^2 \left(1-\alpha_\ell\right)\right)\left(1-\frac{2}{9}\epsilon\right) , \\
P\left(Y_\ell = 2 | X_\ell = 1\right) & = \alpha^3 + \alpha_\ell^2 \left(1-\alpha_\ell\right) + \alpha_\ell \left(1-\alpha_\ell\right)^2 + \left(1-\alpha_\ell\right)^3 ,
\end{array} \right. \\
X_\ell = 2 \left\{ \begin{array}{ll}
P\left(Y_\ell = 0 | X_\ell = 2\right) & = \frac{2}{9}\epsilon, \\
P\left(Y_\ell = 1 | X_\ell = 2\right) & = \frac{4}{9}\epsilon, \\
P\left(Y_\ell = 2 | X_\ell = 2\right) & = \left(1-\frac{6}{9}\epsilon\right) ,
\end{array} \right.
\end{array}
$$

To complete the specification of the HMM we define a prior distribution over the IBD status, $X_\ell$ and a matrix of transition probabilities, $\mathbf{P}_{ij} = P(X_{\ell+1} = j | X_\ell = i)$. The prior distribution is given simply by the expected proportions of IBD status between sibs:

$$
P(X_l = x) = \left\{ \begin{array}{ll}
0.25 & x = 0 \\
0.50 & x = 1 \\
0.25 & x = 2
\end{array} \right.
$$

The transition matrix of the Markov chain depends on the probability, $\rho_\ell$, of a crossover event occurring in either parental meiosis between two consecutive sites, $\ell$ and $\ell + 1$:

$$
\mathbf{P} \;=\; \left( \begin{array}{ccc}
1-\rho_l & \rho_l & 0 \\
\rho_l/2 & 1-\rho_l & \rho_l/2 \\
0 & \rho_l & 1-\rho_l
\end{array} \right)
$$

We set $\rho_\ell = 1 - \exp(-2r_\ell)$ where $r_\ell$ is the genetic distance in Morgans between site $\ell$ and $\ell + 1$ taken from the sex-averaged HapMap genetic map (The International HapMap Consortium, 2005). We assume the probability of more than one crossover event occurring between consecutive markers is negligible.

Using genotypes from chromosome 20 for each pair of putative siblings, we applied the Viterbi algorithm (Rabiner, 1989) to estimate the most likely sequence of IBD states across the chromosome.

# References

Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature*, **437**(7063), 1299–1320.

# 3 Supplementary Tables

| Method | Sample size | Elapsed time (hours) | Memory (GB) | Switch error % |
|---|---:|---:|---:|---:|
| SHAPEIT2 | 1000 | 2.296 | 0.328 | 4.418 |
| SHAPEIT3 | 1000 | 1.946 | 0.489 | 4.477 |
| SHAPEIT3 (4 threads) | 1000 | 0.517 | 0.493 | 4.474 |
| HAPI-UR 1X | 1000 | 0.376 | 1.827 | 19.215 |
| HAPI-UR 3X | 1000 | 1.129 | 1.827 | 14.391 |
| SHAPEIT2 | 2000 | 4.735 | 0.640 | 3.584 |
| SHAPEIT3 | 2000 | 4.086 | 0.970 | 3.725 |
| SHAPEIT3 (4 threads) | 2000 | 1.080 | 0.975 | 3.716 |
| HAPI-UR 1X | 2000 | 0.689 | 2.003 | 13.481 |
| HAPI-UR 3X | 2000 | 2.067 | 2.003 | 9.304 |
| SHAPEIT2 | 5000 | 13.624 | 1.589 | 2.704 |
| SHAPEIT3 | 5000 | 10.926 | 2.430 | 3.019 |
| SHAPEIT3 (4 threads) | 5000 | 2.935 | 2.438 | 3.019 |
| HAPI-UR 1X | 5000 | 1.948 | 3.075 | 6.330 |
| HAPI-UR 3X | 5000 | 5.844 | 3.075 | 4.800 |
| SHAPEIT2 | 10000 | 36.768 | 3.169 | 2.114 |
| SHAPEIT3 | 10000 | 22.691 | 4.851 | 2.544 |
| SHAPEIT3 (4 threads) | 10000 | 6.129 | 4.861 | 2.569 |
| HAPI-UR 1X | 10000 | 6.287 | 5.559 | 4.159 |
| HAPI-UR 3X | 10000 | 18.860 | 5.559 | 3.532 |
| SHAPEIT2 | 20000 | 135.724 | 6.326 | 1.645 |
| SHAPEIT3 | 20000 | 48.218 | 9.691 | 2.108 |
| SHAPEIT3 (4 threads) | 20000 | 12.912 | 9.705 | 2.161 |
| HAPI-UR 1X | 20000 | 18.272 | 10.034 | 3.156 |
| HAPI-UR 3X | 20000 | 54.817 | 10.034 | 2.767 |
| SHAPEIT3 | 49074 | 121.162 | 23.725 | 1.604 |
| SHAPEIT3 (4 threads) | 49074 | 31.887 | 23.755 | 1.609 |
| HAPI-UR 1X | 49074 | 83.254 | 22.028 | 2.241 |
| HAPI-UR 3X | 49074 | 249.763 | 22.028 | 2.059 |

**Supplementary Table 1:** Elapsed time, memory usage and switch-error rate for the UK-BiLEVE data. These results were for chromosome 20 (15,860 SNPs) with switch error calculated on the 384 individuals who had a sibling available for IBD phasing. Each analysis was performed on independent Amazon EC2 m2.2xlarge instances. Time and memory usage were measured using the GNU time command, time is the elapsed real (wall clock) time and memory is the maximum resident set size of the process during its lifetime.

|  | 2Mb | 5Mb | 10Mb |
|---|---|---|---|
| 1,072 | 58.4% | 29.7% | 16.6% |
| 10,072 | 80.2% | 53.1% | 31.8% |
| 152,112 | 95.3% | 85.8% | 68.5% |

**Supplementary Table 2:** The amount of sequence contained in a correctly inferred segments. The table summarizes results of SHAPEIT3 of 3 different runs on the UK Biobank data of different sizes. The estimated haplotypes of the trio children were compared to the trio-based estimates. For each run (row) the percentage of sequence contained in correctly inferred segments of a minimum size (columns) is shown.

| $N$ | Indian | Caribbean |
|---|---|---|
| 1,072 | 10.9% | 22.3% |
| 10,072 | 10.6% | 13.8% |
| 152,112 | 7.3% | 6.5% |

**Supplementary Table 3:** Switch error rates (percentages) in Indian and Caribbean samples. Only one sample of each ancestry was available for switch error estimation, so interpretation of these results should take this in account. Each row represents a dataset of a different size $N$.

| Threads | Walltime (seconds) | × speedup | Efficiency % |
|---:|---:|---:|---:|
| 1 | 3926.25 | 1.00 | 100.00 |
| 2 | 2042.86 | 1.92 | 96.10 |
| 4 | 1095.92 | 3.58 | 89.57 |
| 8 | 609.96 | 6.44 | 80.46 |
| 16 | 345.67 | 11.36 | 70.99 |

**Supplementary Table 4:** SHAPEIT3 timing results for phasing 1,000 samples on chromosome 20 with increasing numbers of threads. Computation was performed on 2×8 Intel Xeon E5-2690 processor with 256 Gb of RAM. Speedup is the ratio of the time over the time for a single-threaded SHAPEIT3 run. Efficiency is the speedup divided by the number of threads (100% efficiency occurring when the speedup is equal to the number of threads used).

| | UK-BiLEVE | UK-BioBank |
|---|---:|---:|
| 0.00<MAF<0.01 | 93240 | 119772 |
| 0.01≤MAF<0.05 | 243207 | 265440 |
| 0.05≤MAF<0.1 | 94228 | 98458 |
| 0.10≤MAF<0.5 | 264061 | 277122 |
| Total | 694736 | 760792 |

**Supplementary Table 5:** Number of SNPs present in the two analysed cohorts.