Supplementary Material

Ancestral Haplotype Reconstruction in Endogamous Populations using Identity-By-Descent

Kelly Finke[1,2], Michael Kourakos[1], Gabriela Brown[1], Yuval B. Simons[3], Alejandro A. Schäffer[4], Rachel L. Kember[5], Maja Bućan[5], Sara Mathieson[6,†]

[1] Department of Computer Science, Swarthmore College, Swarthmore, PA
[2] Department of Biology, Swarthmore College, Swarthmore, PA
[3] Department of Genetics, Stanford University, Stanford, CA
[4] Cancer Data Science Laboratory, National Cancer Institute, NIH, Bethesda, MD
[5] Department of Genetics, University of Pennsylvania, Philadelphia, PA
[6] Department of Computer Science, Haverford College, Haverford, PA
[†] Corresponding author: Sara Mathieson, `smathieson@haverford.edu`

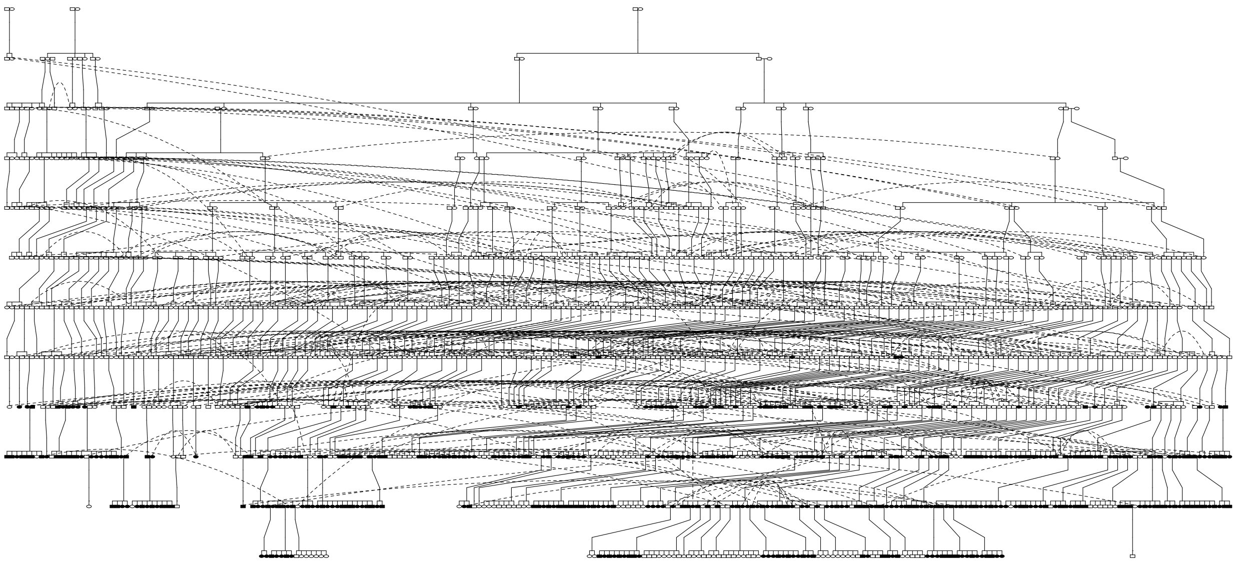

Figure S1: *Pedigree structure: 1338 individuals over 10 generations. Squares represent males and circles represent females. Dotted lines connect the same individual appearing in two different parts of the pedigree. Filled in symbols represent genotyped individuals.*

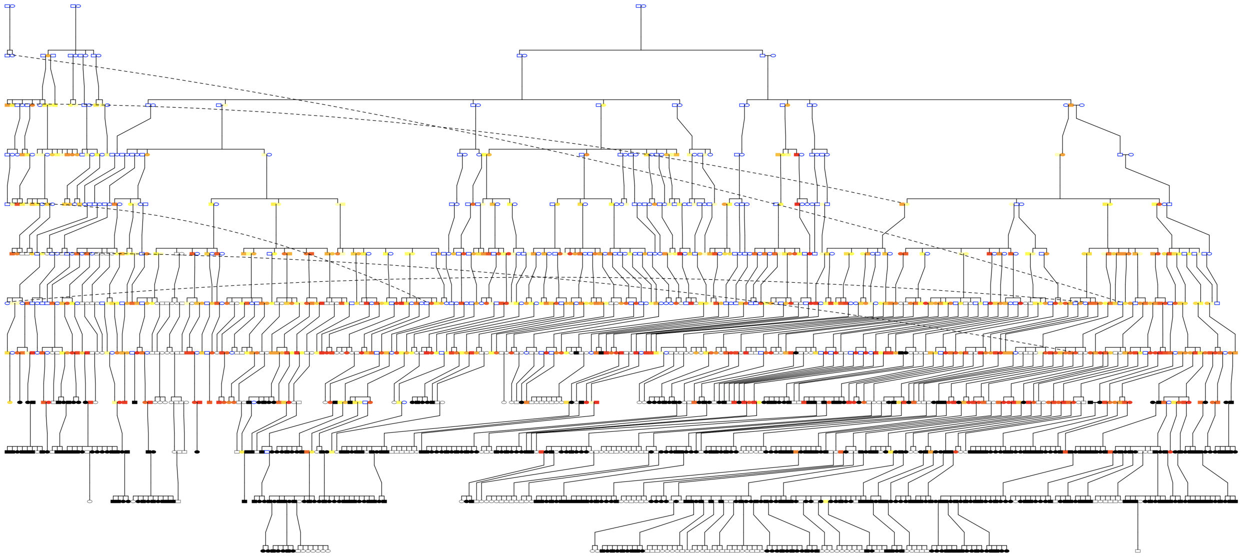

Figure S2: *Position of reconstructed individuals in the pedigree: colors are as follows. Black: genotyped individual, white: no genotyped descendants, yellow-red heatmap: represents number of chromosomes reconstructed, blue: no chromosomes reconstructed. Dotted lines are thinned for clarity, but are the same as in Figure S1.*

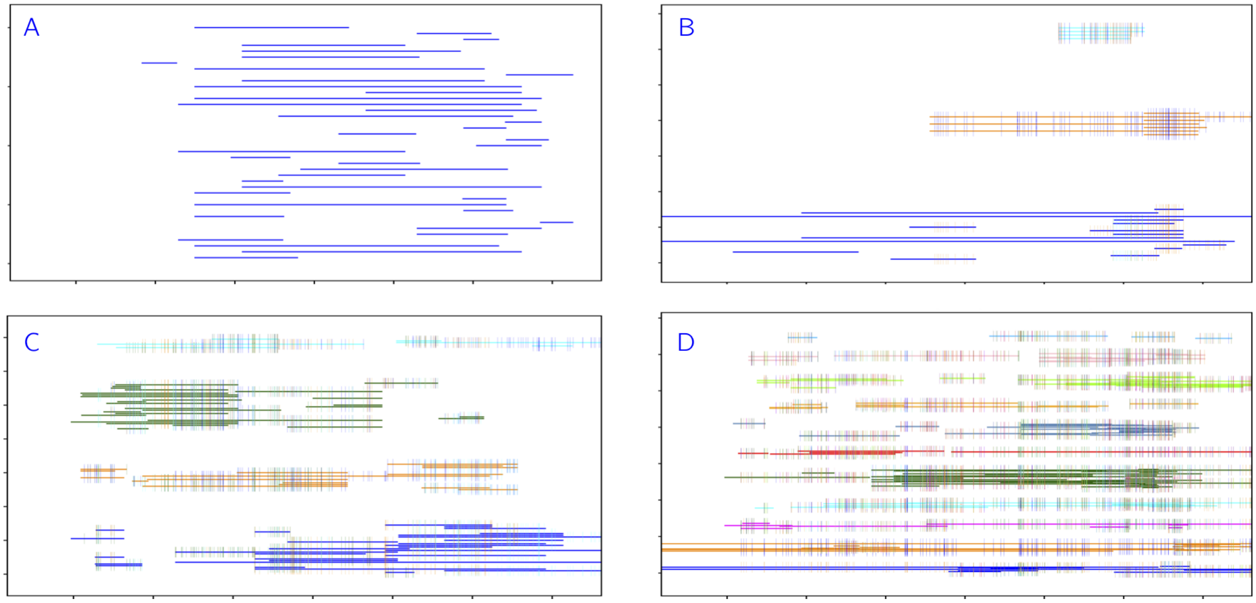

Figure S3: *Unsuccessful reconstruction examples: A) Occasionally we only build one haplotype. B) Sometimes we have a fairly strong reconstruction, but due to the presence of other groups it does not meet our threshold for two strong group. C) Four groups may indicate ambiguity with a spouse or other close relative. D) Sometimes we see many groups and cannot resolve the individual.*

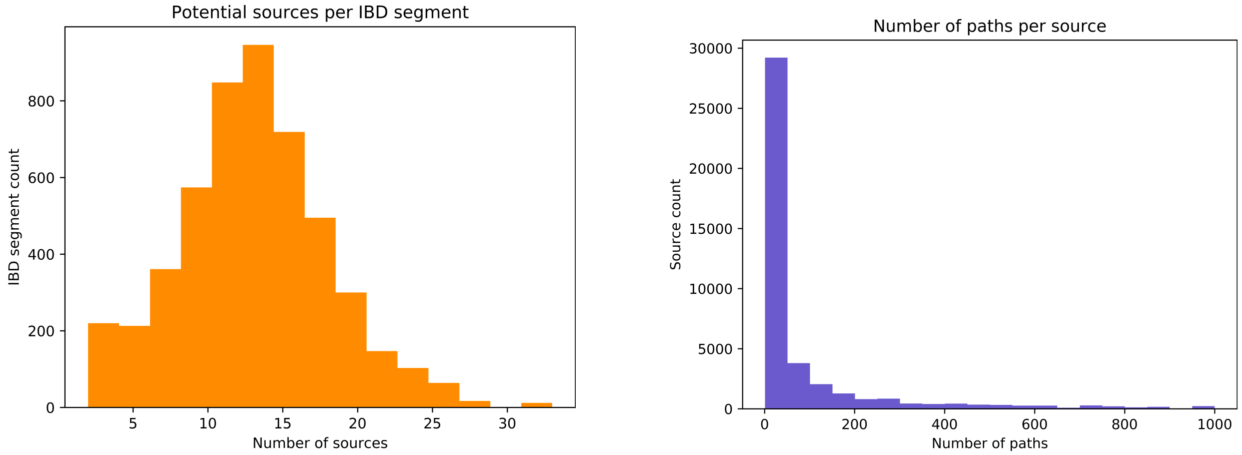

Figure S4: *Source and path distributions for chromosome 21. (left) Distribution of the number of potential sources per IBD segment. (right) Number of paths per source (truncated at 1000, but there is an extremely long tail).*

Algorithm 1: Overview

```
Input: $G$ = genotyped individuals, $NG$ = non-genotyped individuals, $\mathcal{P}$ = pedigree tree
       relating all individuals in $G$ and $NG$
Output: $R$ = reconstructed individuals, $\mathcal{G}_p$ = groups for each individual $p \in R$
find IBDs shared between $G$ using GERMLINE
for $I_k \in IBDs$ do
    $C_k$ = cohort of individuals from $G$ sharing $I_k$
    $S_k$ = sources of $C_k$ (Algorithm 2)
    $d_k(s)$ = number of descendance paths for each $s \in S_k$ (Algorithm 2)
end
$R = G$
$IS$ = list of IBDs to source
while $R$ not changing and $IS$ not empty do
    for $I_k \in IS$ do
        while assignment unsuccessful and $S_k$ is not empty do
            selected source $s^* = \arg\min_s d_k(s)$
            if $d_k(s^*) >$ path threshold then
                ignore $I_k$
            end
            else
                individuals $D_k(s^*)$ = all individuals lying on each path from $s^*$ to $C_k$
                assign $I_k$ to all individuals in $D_k(s^*)$
                if $I_k$ conflicts with reconstructed individual in $D_k(s^*)$ then
                    remove $I_k$ from all $D_k(s^*)$
                    remove $s^*$ from $S_k$
                    assignment round is unsuccessful
                end
            end
        end
    end
    reset $IS$ to empty list
    for individual $p \in NG$ do
        $\mathcal{G}_p$ = reconstructed haplotype groups (Algorithm 3)
        if exactly 2 strong groups in $\mathcal{G}_p$ then
            add $p$ to $R$
        end
        if 2 strong groups and one or more weak groups in $\mathcal{G}_p$ then
            remove weak groups from $\mathcal{G}_p$
            add all IBDs from weak groups to $IS$
            add $p$ to $R$
        end
    end
end
return $R$, $\mathcal{G}_p$ for each $p \in R$
```

Algorithm 2: Source and Descendance Path Finding

```
Input: $C$ = a cohort of individuals sharing a single IBD, $\mathcal{P}$ = pedigree tree containing
       relationships between individuals
Output: $S$ = a list of possible non-redundant sources for cohort $C$
queue $Q$ =list($C$)
for cohort member $p \in C$ do
    multiset $M_p = \{p\}$
end
while $Q$ is not empty do
    individual $p = Q.pop$
    if $p$ is married-in then
        skip the following (married-in have no known ancestors)
    end
    if $p^{(f)}$ has not been processed then
        father's multiset $M_f = M_p$
        father's children set $Ch_f = p$
        add father to $Q$
    end
    else
        extend father's multiset $M_f$ by $M_p$
        add $p$ to father's children set $CH_f$
        add $M_p$ and $p$ to $M$ and $CH$ of any processed ancestors of father
    end
    repeat process for $p^{(m)}$
end
sources $S$ = all individuals $p$ s.t. $M_p$ contains all $c \in C$
for source $s \in S$ do
    $M_{chmax}$ = largest $M_{ch}$ for $ch \in CH_s$
    if length of $M_s = M_{chmax}$ then
        remove redundant source $s$ from $S$
    end
end
for source $s \in S$ do
    if $s.spouse$ in $S$ and $M_s = M_s.spouse$ then
        remove $s$ and $s.spouse$ from $S$
        add couple $s\&s.spouse$ to $S$, s.t. $M = M_s$ and $CH = CH_s$
    end
end
for source $s \in S$ do
    number of descendance paths $d(s) = \prod_{c \in C} m_s(c)$, where $m_s(c)$ = multiplicity of $c$ in $M_s$
end
return $S$ and $d(s)$ for all $s \in S$
```

Algorithm 3: Grouping

```
Input: $R$ = genotyped or reconstructed individuals, $A$ = non-reconstructed individuals,
       ungrouped IBDs $\mathcal{I}_p$ have been placed in each individual $p$
Output: $\mathcal{G}_p$ = groups for each individual $p$
for individual $p \in R$ do
    for IBD $I \in \mathcal{I}_p$ do
        add $I$ to one or both groups in $\mathcal{G}_p$ depending on zygosity
    end
end
for individual $p \in A$ do
    find any homozygous groups $\mathcal{G}_p^{(o)}$
    use overlapping IBDs in $\mathcal{I}_p$ to build heterozygous groups $\mathcal{G}_p^{(e)}$
    duplicate groups in $\mathcal{G}_p^{(o)}$ and create $\mathcal{G}_p = \mathcal{G}_p^{(o)} \cup \mathcal{G}_p^{(e)}$
    remove all IBDs from $\mathcal{S}_p$ that were used to build groups in $\mathcal{G}_p$
    for pairs of groups $G_i, G_j \in \mathcal{G}_p$ and remaining IBD $I \in \mathcal{I}_p$ do
        if $I$ overlaps $G_i$ and $G_j$ sufficently then
            merge $G_j$ into $G_i$ and delete $G_j$
        end
    end
    for pairs of groups $G_i, G_j \in \mathcal{G}_p$ do
        if $G_i$ and $G_j$ overlap or "line up" then
            merge $G_j$ into $G_i$ and delete $G_j$
        end
    end
end
```