Supplementary Methods and Figures

October 22, 2015

Contents

1 Algorithm Pseudocode

1.1 Algorithm 1 - MakeCopySet

Algorithm 1 MakeCopySet - ‘Rare variant copy state selection’

```
Require: $\boldsymbol{H}$                                   ▷ A panel of reference haplotypes
Require: $\boldsymbol{G}$                     ▷ A list of genotypes for the sample to be phased
Require: $\boldsymbol{C}$                  ▷ A list of allele counts of sites common to G and H
Require: W                                                   ▷ The width of a window
Require: K                                 ▷ The number of copying states in each window
Require: M                                         ▷ The total number of panel haplotypes
Require: pos ()                                ▷ pos $(i)$ returns the position of the i$^{\mathrm{th}}$ site

procedure MakeCopySet($\boldsymbol{H}, \boldsymbol{G}, \boldsymbol{C}, W, K, M$)

    create empty lists Sites [ ] and CopyStates [ ]
    $\boldsymbol{B} \leftarrow$ MakeWindowBoundaries(W)
    $j \leftarrow 1$
    for $i \leftarrow 1,$ length $(\boldsymbol{G})$ do
        if pos $(i) > \boldsymbol{B}[j+1]$ then
            $j \leftarrow j+1$
        if $(\boldsymbol{G}[i] \geq 1) \wedge (\boldsymbol{C}[i] \geq 1)$ then
            add $(i, \boldsymbol{C}[i])$ to Sites $[j]$
    for $j \leftarrow 1,$ length (Sites) do
        sort (Sites $[j]$)                                         ▷ Sort by allele count
    for $j \leftarrow 1,$ length (Sites) do
        if length (Sites $[j]$) $== 0$ then
            $w \leftarrow$ index of nearest window such that length (Sites $[w]$) $> 0$
            add Sites $[w]$ to Sites $[j]$
    $k \leftarrow 0$
    for $j \leftarrow 1,$ length $(\boldsymbol{B}) - 1$ do
        $m \leftarrow 1$
        while $m \leq M \wedge k < K$ do
            CopyStates $[j] \leftarrow$ AddStates(CopyStates $[j]$, Sites $[j]$, $\boldsymbol{H}, m, K-k$)
            $k \leftarrow$ length (CopyStates $[j]$)
        $m \leftarrow m+1$
```

1.2 Algorithm 2 - AddStates

Algorithm 2 AddStates - 'add copy states matching $\boldsymbol{G}$ at sites with allele count m'

```
function AddStates(CopyStates, Sites, $\boldsymbol{H}$, $m$, $r$)
    create empty list States
    $i \leftarrow 1$
    $(j, c) \leftarrow$ Sites$[i]$                    ▷ $j$ is the site index, $c$ is the allele count
    while $c \leq m$ do
        $i \leftarrow i + 1$
        $(j, c) \leftarrow$ Sites$[i]$
        if $c == m$ then
            $h \leftarrow 1$
            while $h \leq$ length$(\boldsymbol{H}[j])$ do                    ▷ Scan $j^{th}$ row of $\boldsymbol{H}$
                if $\boldsymbol{H}[j, h] == 1$ then
                    if $h \notin$ CopyStates then
                        add $h$ to States
    if length(States) $> r$ then
        remove length(States) $- r$ randomly chosen elements
    for all $h \in$ States do
        if $h$ is odd then
            add $h$ and $h + 1$ to States
        else
            add $h$ and $h - 1$ to States
    add States to CopyStates
```

2 Supplementary Figures

2.1 Supplementary figure 1

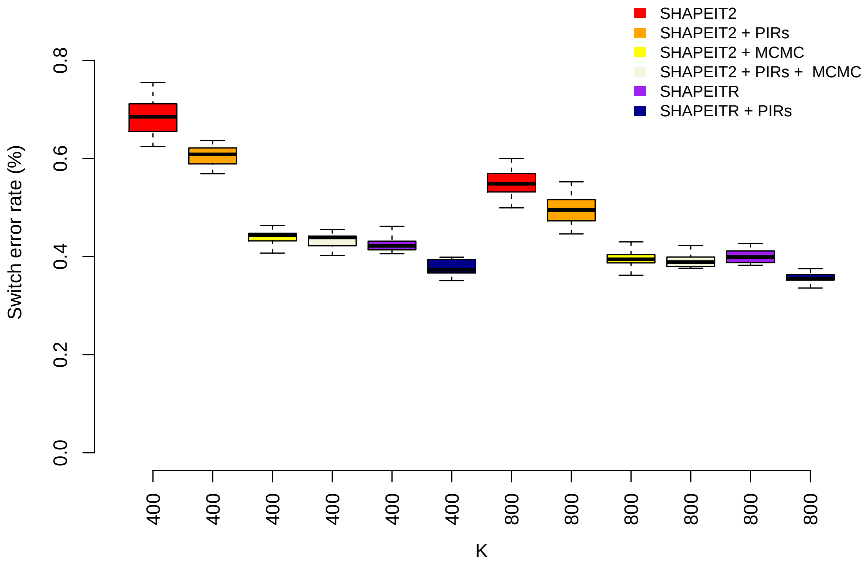

Figure 1: **Comparison of switch error rates for trio parents using half of the original UK10K panel (3756 haplotypes).** This figure corresponds to figure 1 in the main text but with a reduced reference panel consisting of half of the original UK10K panel (3756 haplotypes). The box-plot compares the empirical distribution of switch error rates achieved by different methods in 20 different phasing runs of chromosome 20 averaged over the two trio parents. Two different numbers of copying states were used: $K \in \{400, 800\}$ for a single, fixed window size of 0.5Mb. Methods compared are SHAPTEITR and SHAPTEIT2 with and without use of MCMC.

2.2 Supplementary figure 2

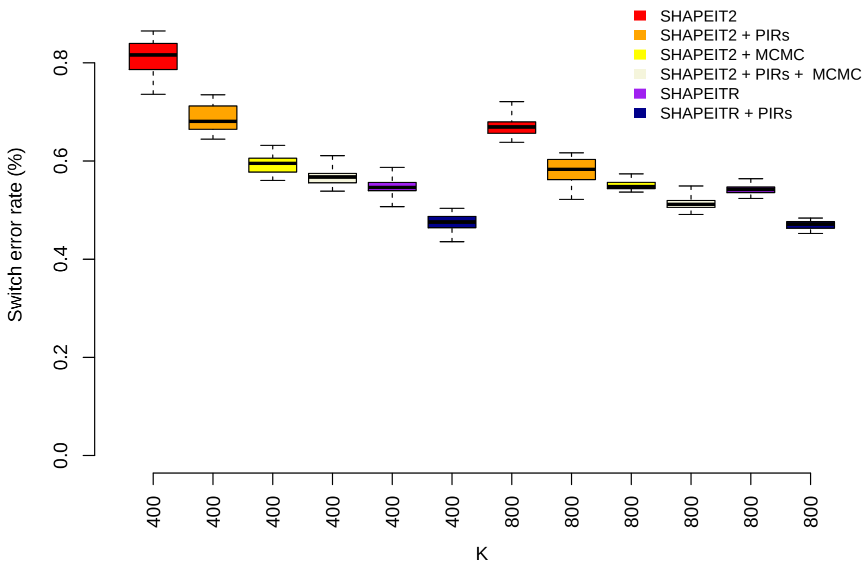

Figure 2: **Comparison of switch error rates for trio parents using one quarter of the original UK10K panel (1878 haplotypes).** This figure corresponds to figure 1 in the main text but with a reduced reference panel consisting of one quarter of the original UK10K panel (1878 haplotypes) The box-plot compares the empirical distribution of switch error rates achieved by different methods in 20 different phasing runs of chromosome 20 averaged over the two trio parents. Two different numbers of copying states were used: $K \in \{400, 800\}$ for a single, fixed window size of 0.5Mb. Methods compared are SHAPTEITR and SHAPTEIT2 with and without use of MCMC.

2.3 Supplementary figure 3

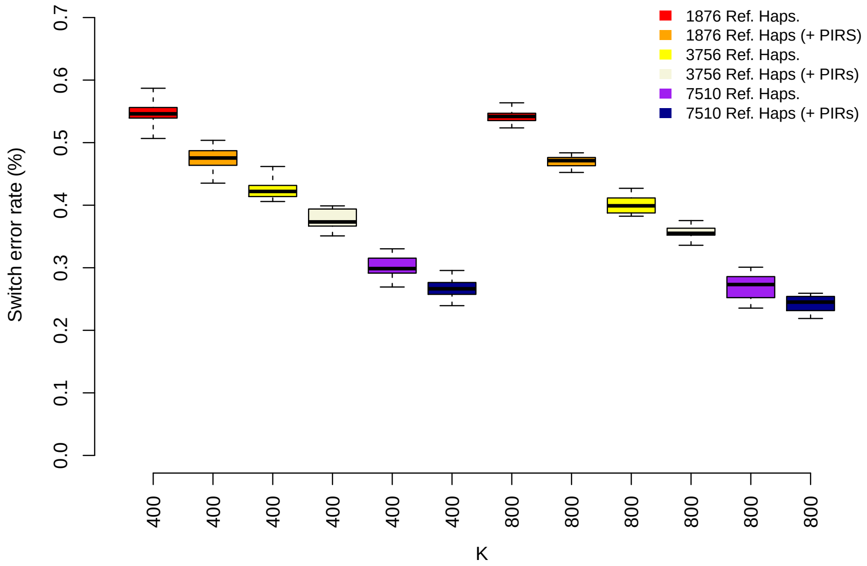

Figure 3: **Variation of switch error rates with panel size** The boxplot compares the empirical distribution of switch error rates achieved by SHAPEITR using 3 different panel sizes - 7510, 3756 and 1878 haplotypes sampled from the UK10K panel. Boxes show the distribution of errors in 20 different phasing runs of chromosome 20 averaged over the two trio parents. Two different numbers of copying states were used: $K \in \{400, 800\}$ for a single, fixed window size of 0.5Mb.

2.4 Supplementary figure 4

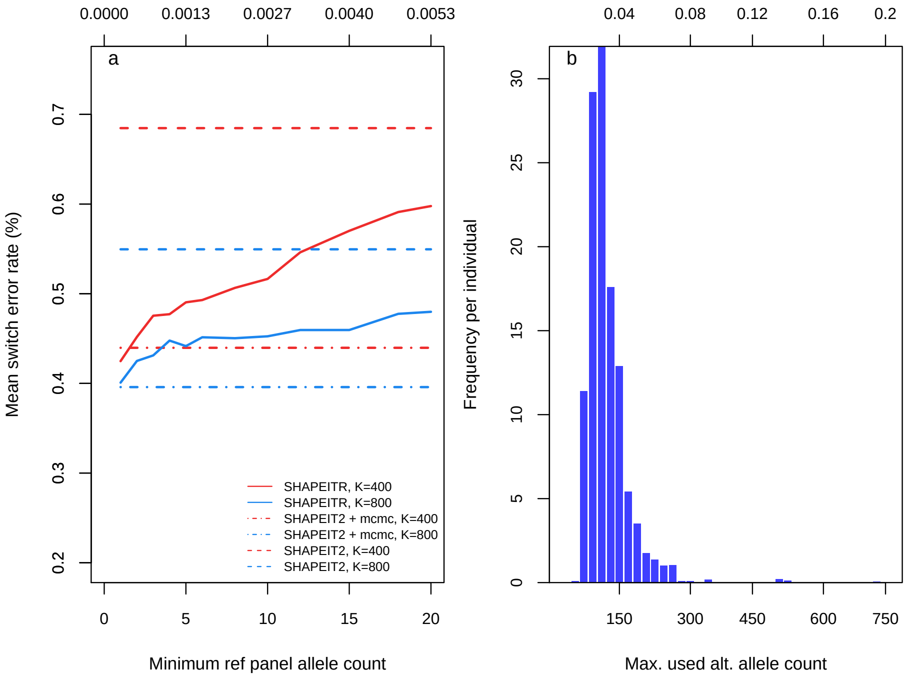

Figure 4: **Properties of using rare variants for state selection with a smaller panel (3756 haplotypes)**. This figure **Fig 4a** Effect on switch error rate of varying the minimum minor allele count used for selecting individuals from whom to copy in SHAPEITR. Horizontal axes: minimum minor allele count (bottom) and corresponding frequency in panel (top) used for selection. Solid lines: mean switch error rates for SHAPEITR; dashed (and dashdot) lines: mean switch error rates for SHAPEIT2 with (and without) MCMC. Colours indicate whether K = 400 (red) or K = 800 (blue) copying states were used. In both cases, errors refer to phasing the whole of chromosome 20 and were averaged over both trio parents and 20 runs. **Fig 4b** Distribution of maximum allele counts used for matching in a single window when choosing $K = 400$ copying states. Horizontal axis: maximum minor allele count (bottom) and corresponding frequency in the reference panel (top) of a site used for matching. Vertical axis: frequency averaged over both trio parents and 20 different runs. Each bar represents a bin of width $\simeq 0.0027$ corresponding to an allele count of 20.