Supplemental Data

Haplotype Estimation Using Sequencing Reads

Olivier Delaneau, Bryan Howie, Anthony J. Cox, Jean-Francois Zagury, and Jonathan Marchini

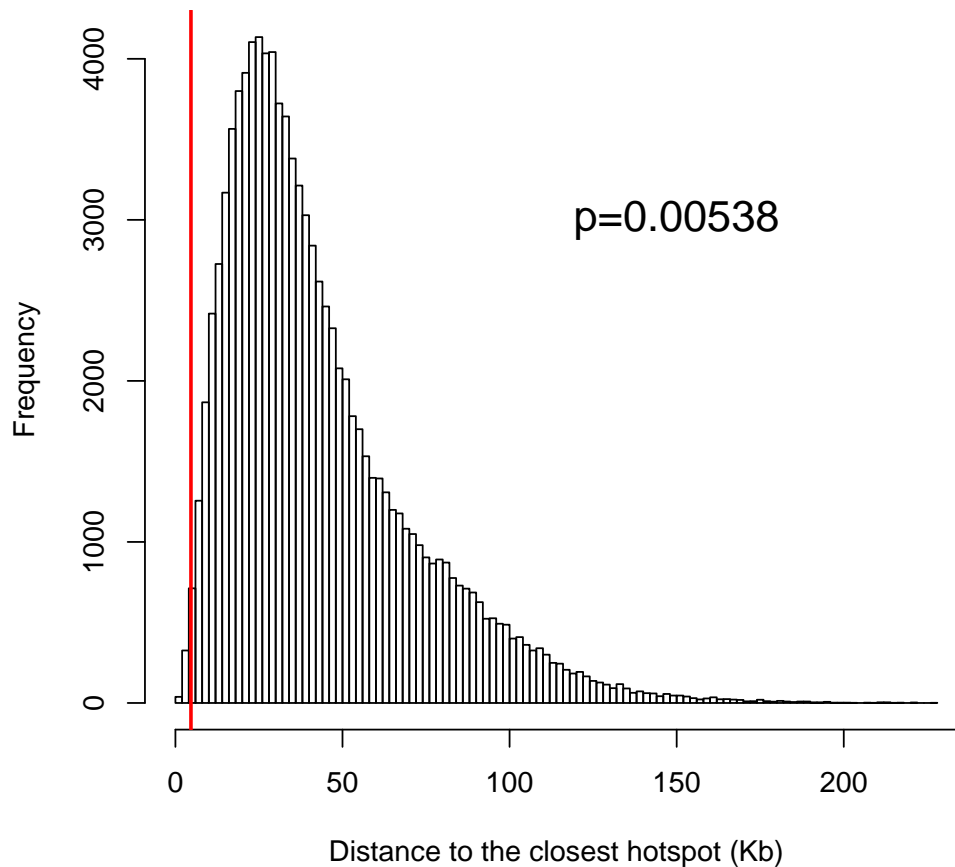## Supplemental Inventory

Supplemental Figures and Tables

Supplemental Materials

Figure S1 : Distribution of the distances between 100,000 sets of 4 ran-domly chosen physical positions on chromosome 20 and their closest respective hotspots. 100,000 sets of 4 physical positions were randomly drawn in the reachable position range of chromosome 20. For each set, we measured the distances between the 4 random positions to their closest flanking hotspots (recombination rate $\geq$ 10 cM/Mb) and average them. The 100,000 mean distances were then plotted, resulting in this distribution of distances. Mean distance for the 4 singletons incorrectly phased is shown with the red vertical line, leading to an empirical p-value of 0.00538.

Software	TGP phased	TGP coverage	Trio coverage	Running time (h)
Beagle	no	0x	0x	3.95
SHAPEIT2	no	0x	0x	20.6
SHAPEIT2	no	0x	5x	20.7
SHAPEIT2	no	0x	10x	20.6
SHAPEIT2	no	0x	20x	20.8
SHAPEIT2	no	4x	0x	38.9
SHAPEIT2	no	4x	5x	39
SHAPEIT2	no	4x	10x	38.3
SHAPEIT2	no	4x	20x	38.5
SHAPEIT2	yes	0x	0x	0.1
SHAPEIT2	yes	0x	5x	0.2
SHAPEIT2	yes	0x	10x	0.3
SHAPEIT2	yes	0x	20x	0.5

Table S1 : Comparison of running times on the Illumina trio parents and TGP samples. The second column (TGP phased) indicates if the haplotypes of 379 TGP European samples are considered as phased (yes) or have to be phased jointly with the Illumina trio parents (no). The third column (TGP coverage) tells if any reads were used to phase the TGP European samples. The forth column (Trio coverage) shows the read coverage used to phase the Illumina trio parents. All the running times measured for SHAPEIT2 combine both the times needed to parse the BAM files and to run the MCMC algorithm. SHAPEIT2 was run using 1 thread.