

Electronic Supplementary Material 1. Population samples used in this study. Code is the code used by the 1000 genomes project, N is the number of individuals processed in the current study. Almost all available individuals were analysed, the exceptions being a small number of unusually small file sizes and failed run.

Code	population	region	N
GWD	Gambia	Africa	109
ESN	Esan	Africa	85
MSL	Mende	Africa	79
YRI	Yoruba	Africa	51
LWK	Luhya	Africa	31
BEB	Bengal	CS Asia	75
PJL	Punjab	CS Asia	43
STU	Sri Lanka	CS Asia	94
ITU	Indian Telugu	CS Asia	86
CHS	Southern Han	E Asia	40
KHV	Korea	E Asia	47
CDX	Chinese Dai	E Asia	31
CHB	Han, Beijing	E Asia	70
JPT	Japan	E Asia	35
GBR	Britain	Europe	40
FIN	Finland	Europe	31
IBS	Spain	Europe	51
TSI	Italy	Europe	111