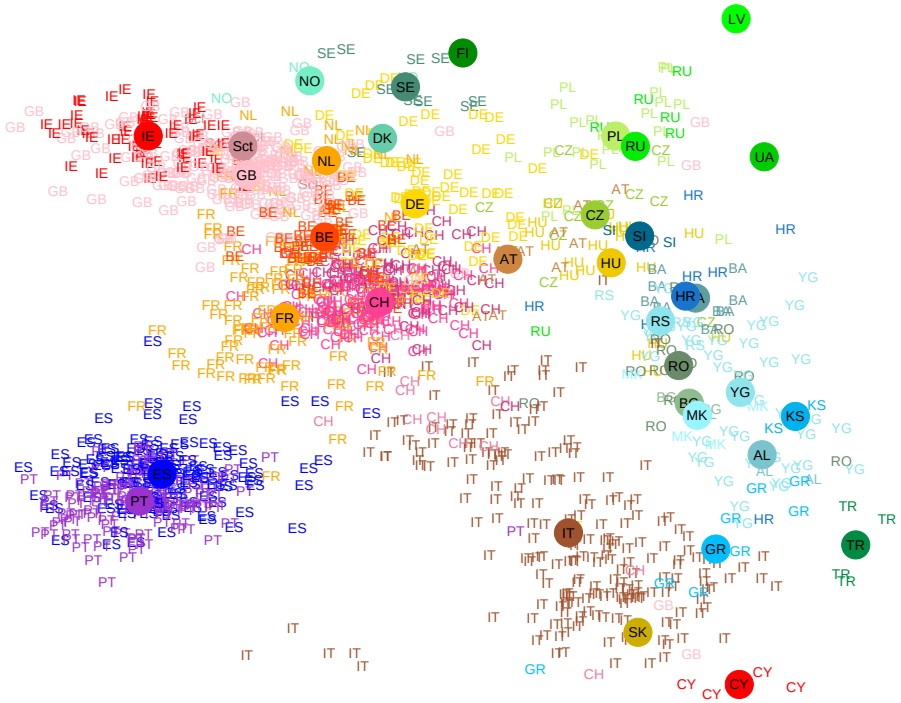
Supplementary Information for: A model based approach for analysis of spatial structure in genetic data

Wen-Yun Yang[1,2], John Novembre[1,3], Eleazar Eskin[1,2,4,7,8], and Eran Halperin[5,6,7]

[1]Interdepartmental Program in Bioinformatics,
[2]Department of Computer Science,
[3]Department of Ecology and Evolutionary Biology and
[4]Department of Human Genetics, University of California, Los Angeles, California, USA
[5]International Computer Science Institute, Berkeley, California, USA
[6]School of Computer Science and the Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel
[7]These authors contribute equally to this work
[8]Correspondence should be addressed to E.E. (eeskin@cs.ucla.edu)

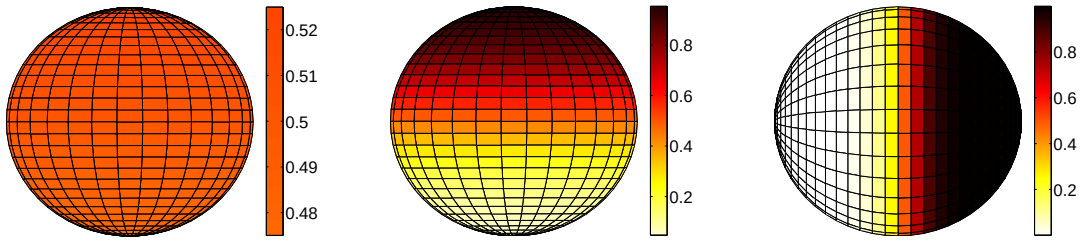
Nature Genetics: doi:10.1038/ng.2285

Supplementary Figures



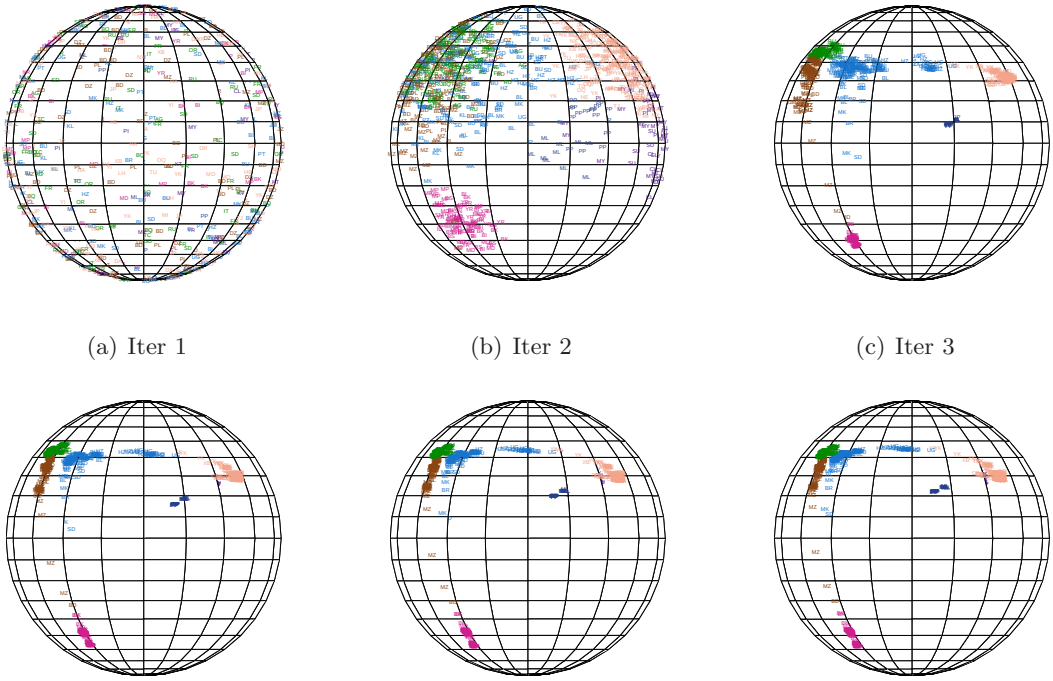
Supplementary Figure 1: Mapping results on POPRES data set by placing individuals using country of origin information. A 10-fold cross validation is performed. In each run, we fit the slope function using the true location information of 90% of the individuals and predict the location for the remaining 10%.

Nature Genetics: doi:10.1038/ng.2285



(a) Tiny change (b) Gradual change (c) Sharp change in different direction

Supplementary Figure 2: Examples of the allele frequency model for a sphere. The allele frequency is represented by different colors (yellow are low allele frequencies, while red are high allele frequencies). (a) A SNP with constant allele frequency over the sphere. (b) A SNP with graduate allele frequency changes over the sphere. (c) A SNP with sharp frequency changes. The parameter a is set to $[0, 0, 0.1]$, $[0, 0, 3]$ and $[10, 0, 0]$, respectively and the parameter b is set to be zero in all three spheres. The sphere coordinates are drawn from a unit sphere, i.e. $||\vec{x}|| = 1$.

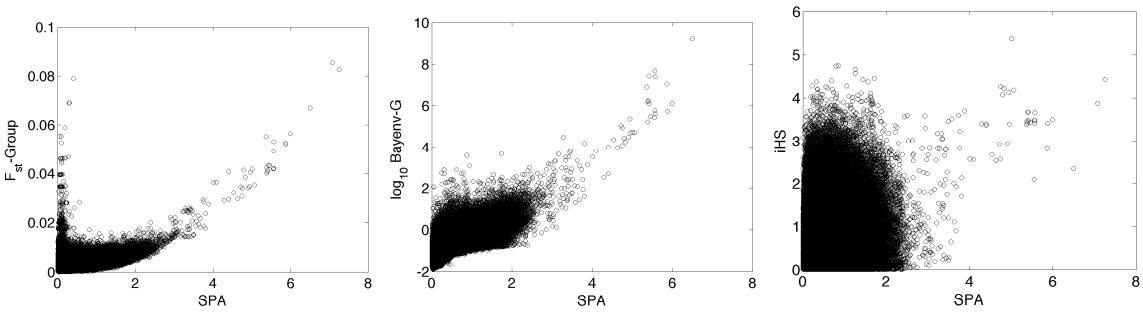


(a) Iter 1 (b) Iter 2 (c) Iter 3

(d) Iter 4 (e) Iter 5 (f) Iter 6

Supplementary Figure 3: Globe mapping convergence with random initialization for individuals from the HGDP data set. The colors represent the continent-level origins for each individual. Iteration 1 starts from random positioning of individuals. By iteration 4, the algorithm separates the continents.

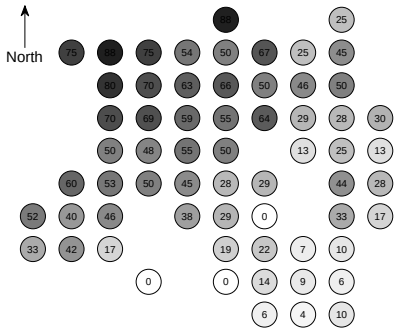
Nature Genetics: doi:10.1038/ng.2285



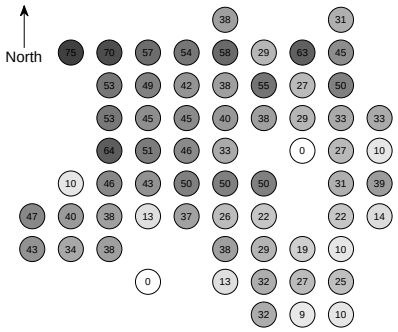
(a) SPA and F_{ST} (b) SPA and Bayenv (c) SPA and iHS

Supplementary Figure 4: Correlation between the SPA and other methods. F_{ST} and Bayenv use geographical groups to divide populations. The environmental factor in Bayenv is the individual coordinates in the first principle component.

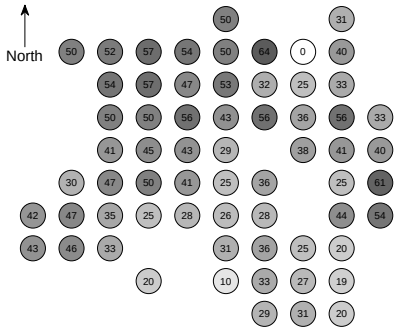
Nature Genetics: doi:10.1038/ng.2285



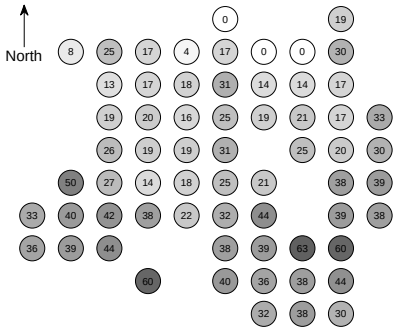
(a) *LCT*: rs6730157



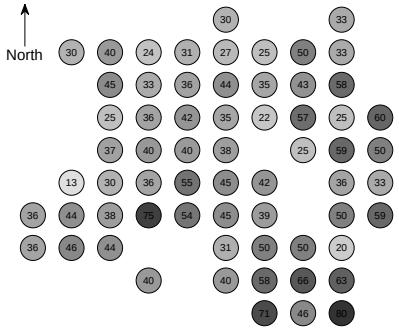
(b) *HLA*: rs9268560



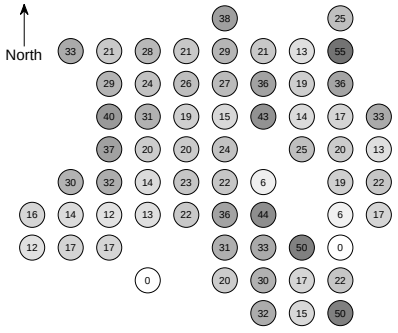
(c) *FOXP2*: rs2106900



(d) *OCA2*: rs916977



(e) *LRP1B*: rs7598314



(f) Typical SNP

Supplementary Figure 5: Spatial distribution of SNPs with extreme allele frequency gradients. The grey scale stands for allele frequency: dark for high frequency and white for low frequency. The number in the circle stands for allele frequency percentage values. We divide the whole map into 10×10 grid. We then calculate the allele frequency for each small region by averaging all individuals in the region. Regions with less than 5 individuals are removed for accurate allele frequency estimation.

Nature Genetics: doi:10.1038/ng.2285

Supplementary Tables

Supplementary Table 1: Summary of SPA globe mapping results. The mean and standard deviation of each continent population are calculated based on an alignment to actual world globe. The positive and negative latitudes stand for north and south latitudes, respectively. The positive and negative longitudes stand for east and west longitude, respectively. The longitude and latitude for each continent are computed by averaging the individual spatial assignments.

Continent	Pred. Latitude	Pred. Longitude	Actual Latitude	Actual Longitude
Africa	-44.005 ± 4.030	19.548 ± 0.885	1.845 ± 10.977	12.634 ± 16.229
America	21.206 ± 1.988	-151.095 ± 3.449	9.344 ± 16.035	-82.453 ± 18.161
Central South Asia	37.628 ± 3.625	28.047 ± 11.212	32.051 ± 4.722	69.550 ± 3.979
East Asia	30.683 ± 2.617	95.598 ± 4.925	36.685 ± 12.841	115.542 ± 14.250
Europe	38.582 ± 1.434	10.630 ± 2.751	47.732 ± 7.618	13.592 ± 16.317
Middle East	29.799 ± 7.600	12.581 ± 1.277	31.718 ± 0.451	29.307 ± 12.276
Oceania	24.203 ± 1.007	66.530 ± 2.189	-4.741 ± 0.984	147.444 ± 5.905

Supplementary Table 2: Correlation coefficient between six methods. We check whether a SNP with top 2% scores for POPRES set or 1% scores for HapMap set in 100kb window to determine whether this window is under selection. Then we have a binary vector across the whole genome for all four methods, and we compute the correlation coefficient using the binary vector.

	iHS	F_{ST}-Grp	F_{ST}-Cntry	Bayenv-Grp	Bayenv-Cntry	SPA
iHS	1.000000					
F_{ST}-Grp	0.028713	1.000000				
F_{ST}-Cntry	0.013455	0.401576	1.000000			
Bayenv-Grp	0.018016	0.305782	0.123768	1.000000		
Bayenv-Cntry	0.019460	0.203627	0.103108	0.333595	1.000000	
SPA	0.023986	0.333148	0.123805	0.446488	0.296796	1.000000

Nature Genetics: doi:10.1038/ng.2285

Supplementary Table 3: Genes with extreme gradients detected only by SPA.

Genes	SNP with extreme gradient	highest SPA score	Position
MGLL	rs782437	2.88928	chr3:128899892
SCAND3,ZNF192,ZSCAN	rs6903535	2.74179	chr6:28525201
FOXP2	rs2106900	2.71077	chr7:113909742
TMEM104	rs2385067	2.57703	chr17:70321665
ENY2,NUDCD1	rs1380098	2.54717	chr8:110317808
TAS2R,PRH1,PRR4	rs2597996	2.54539	chr12:11099651
MIR1244,BCL2L14	rs4763782	2.51019	chr12:12146023
ITPR1	rs7637793	2.49356	chr3:4725558
NUP153	rs11753865	2.48271	chr6:17790701
ZAP70	rs6736735	2.48015	chr2:97702865
TMEM117	rs2407790	2.46361	chr12:42717760
ERC1	rs11061714	2.44995	chr12:1348831
SEMA6D	rs281297	2.44008	chr15:45472796
PTK2B,DPYSL2,TRIM35	rs6557991	2.42558	chr8:27231225
SLC45A1	rs1466654	2.42438	chr1:8298500
SLC24A3	rs4814838	2.40770	chr20:19267846
SOX6	rs7118395	2.40590	chr11:16155641
SPOCK1	rs2348605	2.39459	chr5:136838003
ADAMTSL3	rs7169595	2.38401	chr15:82196578
WWOX	rs441004	2.38284	chr16:77794331
SLC26A4,LOC286002	rs11769313	2.38096	chr7:107101183
NRXN3	rs11625485	2.36487	chr14:79311885
ZNF19	rs2288486	2.35797	chr16:70070535
SEMA3E	rs215302	2.35116	chr7:83051200
ZDHHC2	rs2959634	2.35056	chr8:17071223
CDH7	rs7237421	2.35029	chr18:61586240
KCNIP4	rs7689421	2.33900	chr4:20364338
AK5	rs11162351	2.33449	chr1:77717320
ZAK	rs3754744	2.33022	chr2:173710310
IDH2	rs12443387	2.32722	chr15:88453860
USH2A	rs2677112	2.32663	chr1:213891790
FLJ22536	rs2078482	2.32446	chr6:22105905
ERC2	rs7628951	2.32417	chr3:56432745
PRRX1	rs593479	2.32273	chr1:168909523
RBFOX3	rs4790055	2.31235	chr17:75001008
FUT11,SEC24C	rs3849969	2.30577	chr10:75196005
GRRP1	rs1335759	2.28446	chr1:26344330
SORBS1	rs526928	2.28296	chr10:97324281
NEBL	rs1340293	2.28250	chr10:21239171
HUNK	rs2833609	2.27600	chr21:32303489
TMEM170B	rs9469574	2.27332	chr6:11698126

Nature Genetics: doi:10.1038/ng.2285

Supplementary Table 4: A full list of regions with extreme gradients detected by SPA. The 450 (0.1% of total) SNPs with the highest scores are listed. Results of different SNPs are merged if the two SNPs are at distance smaller than 1MB.

Genes	SNP with most extreme gradient	highest SPA score	Position
LCT region	rs6730157	7.25764	chr2:135623558
HLA-DPB1 region	rs9268560	3.67652	chr6:32497490
HLA-B region	rs2517510	3.50633	chr6:31138101
HERC2,OCA2	rs916977	3.42020	chr15:26186959
ADH1C	rs1789903	3.07810	chr4:100481064
LOC283177	rs4592433	3.06241	chr11:133853131
FAM114A1,TLR10,TLR1	rs6835514	3.02649	chr4:38570775
MGLL	rs782437	2.88928	chr3:128899892
PLA2R1,LY75, LY75-CD302,ITGB6	rs16844715	2.83868	chr2:160623352
BNC2	rs10756762	2.81414	chr9:16553123
HPS5,GTF2H1	rs4150581	2.76205	chr11:18313846
SCAND3,ZNF192, ZSCAN16,ZSCAN23	rs6903535	2.74179	chr6:28525201
TWSG1	rs8091539	2.73925	chr18:9398889
FOXP2	rs2106900	2.71077	chr7:113909742
SCN2A,SCN1A	rs1461197	2.67766	chr2:166632718
RFPL1,RFPL1S	rs5763240	2.65562	chr22:28166926
LRP1B	rs7598314	2.63123	chr2:142250435
SYT1	rs7308297	2.62652	chr12:78301975
TMEM104	rs2385067	2.57703	chr17:70321665
ZFAND3	rs10485029	2.56609	chr6:37907682
ENY2,NUDCD1	rs1380098	2.54717	chr8:110317808
TAS2R50,TAS2R19, TAS2R31,TAS2R20, PRH1,PRR4,TAS2R46	rs2597996	2.54539	chr12:11099651
MIR1244,BCL2L14,LRP6	rs4763782	2.51019	chr12:12146023
SUCLG2	rs1352657	2.50426	chr3:67535708
CHMP1A,DPEP1,C16orf55	rs164749	2.50165	chr16:88235725
ITPR1	rs7637793	2.49356	chr3:4725558
NUP153	rs11753865	2.48271	chr6:17790701
ZAP70	rs6736735	2.48015	chr2:97702865
EHBP1,OTX1	rs11125946	2.47946	chr2:63151654
TMEM117	rs2407790	2.46361	chr12:42717760
RBFOX1	rs11645481	2.45097	chr16:7021094
ERC1	rs11061714	2.44995	chr12:1348831

Continued on Next Page...

Nature Genetics: doi:10.1038/ng.2285

Genes	SNP with most extreme gradient	highest SPA score	Position
ACOT6,ACOT4	rs4903128	2.44437	chr14:73145688
SEMA6D	rs281297	2.44008	chr15:45472796
PTK2B,DPYSL2,TRIM35	rs6557991	2.42558	chr8:27231225
LOC729234	rs1917890	2.42547	chr2:96035728
SLC45A1	rs1466654	2.42438	chr1:8298500
SLC24A3	rs4814838	2.40770	chr20:19267846
SOX6	rs7118395	2.40590	chr11:16155641
LOC732275	rs8051237	2.40286	chr16:84939020
LRRFIP1	rs6754972	2.39608	chr2:238206505
SPOCK1	rs2348605	2.39459	chr5:136838003
ADAMTSL3	rs7169595	2.38401	chr15:82196578
WWOX	rs441004	2.38284	chr16:77794331
SLC26A4,LOC286002	rs11769313	2.38096	chr7:107101183
EMILIN2	rs592120	2.38076	chr18:2875118
VAV3	rs10494081	2.37404	chr1:108203626
NRXN3	rs11625485	2.36487	chr14:79311885
ZNF19	rs2288486	2.35797	chr16:70070535
SEMA3E	rs215302	2.35116	chr7:83051200
ZDHHC2	rs2959634	2.35056	chr8:17071223
CDH7	rs7237421	2.35029	chr18:61586240
FOXN3	rs1952182	2.34064	chr14:88978953
KCNIP4	rs7689421	2.33900	chr4:20364338
AK5	rs11162351	2.33449	chr1:77717320
ZAK	rs3754744	2.33022	chr2:173710310
IDH2	rs12443387	2.32722	chr15:88453860
USH2A	rs2677112	2.32663	chr1:213891790
FLJ22536	rs2078482	2.32446	chr6:22105905
ERC2	rs7628951	2.32417	chr3:56432745
ITGAX	rs1106398	2.32299	chr16:31277953
PRRX1	rs593479	2.32273	chr1:168909523
RBFOX3	rs4790055	2.31235	chr17:75001008
VAT1L	rs33967759	2.30990	chr16:76546435
FUT11,SEC24C	rs3849969	2.30577	chr10:75196005
UGT2B11	rs6817250	2.28873	chr4:70123190
GRRP1	rs1335759	2.28446	chr1:26344330
SORBS1	rs526928	2.28296	chr10:97324281
NEBL	rs1340293	2.28250	chr10:21239171
LOC100188947	rs12246543	2.27656	chr10:93291408
HUNK	rs2833609	2.27600	chr21:32303489
TMEM170B	rs9469574	2.27332	chr6:11698126

Nature Genetics: doi:10.1038/ng.2285

Supplementary Notes

Newton's method for optimizing SPA likelihood function

Newton's method is a widely known algorithm for minimizing a convex function. In each iteration, it needs the first and second derivatives to determine the search direction. For x_i in (3), the first and second derivatives can be efficiently calculated as follows

$$\begin{aligned}\frac{\partial L}{\partial x_i} &= -\sum_j \left[g_{ij}(1-f_{ij}) - (2-g_{ij})f_{ij}\right] \cdot a_j \\ \frac{\partial^2 L}{\partial x_i^2} &= \sum_j 2f_{ij}(1-f_{ij})a_j a_j^T\end{aligned}$$

Similarly, for a_j and b_j in (4) those can be calculated as follows

$$\begin{aligned}\frac{\partial L}{\partial a_j} &= -\sum_i \left[g_{ij}(1-f_{ij}) - (2-g_{ij})f_{ij}\right] \cdot x_i \\ \frac{\partial L}{\partial b_j} &= -\sum_i \left[g_{ij}(1-f_{ij}) - (2-g_{ij})f_{ij}\right]\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial a_j^2} &= \sum_i 2f_{ij}(1-f_{ij})x_i x_i^T \\ \frac{\partial^2 L}{\partial b_j^2} &= \sum_i 2f_{ij}(1-f_{ij}).\end{aligned}$$

The computational complexity for each iteration of this algorithm is $O(NLK^2)$. The total computational time depends on the number of iterations for the algorithm to converge. In the data sets we analyzed in this paper, we used 10 to 20 iterations.

Pseudo-Newton's method for admixed individual positioning

To achieve fast convergence for the admixed individual positioning problem, we again use Newton's method to optimize the log-likelihood function (5). The first and second derivatives in x and y up to a constant can be computed as follows:

$$\begin{aligned}\frac{\partial L}{\partial x} &= \sum_j \left[I(g_j=2)(1-p_j) + I(g_j=1)t_1 + I(g_j=0)(-p_j)\right] a_j \\ \frac{\partial^2 L}{\partial x^2} &= \sum_j \left[I(g_j=2)(1-p_j)p_j + I(g_j=1)t_2 + I(g_j=0)(1-p_j)p_j\right] (-a_j a_j^T) \\ \frac{\partial^2 L}{\partial x \partial y} &= \sum_j I(g_j=1)\left[\frac{m_j(1-m_j)(1-2m_j)p_j(1-p_j)(1-2p_j)}{\left[(1-m_j)p_j + (1-p_j)m_j\right]^2} + \frac{2m_j(1-m_j)p_j(1-p_j)}{(1-m_j)p_j + (1-p_j)m_j}\right](-a_j a_j^T)\end{aligned}$$

Nature Genetics: doi:10.1038/ng.2285

where I is an indicator function equal to one if the condition holds and zero otherwise, and

$$\begin{aligned} t_1 &= \frac{(1-2m_j)(1-p_j)p_j}{p_j(1-m_j)+m_j(1-p_j)} \\ t_2 &= (1-2m_j)\frac{\frac{(1-m_j)p_j}{1-p_j} - \frac{m_j(1-p_j)}{p_j}}{\left(\frac{1-m_j}{1-p_j}+\frac{m_j}{p_j}\right)^2} \end{aligned}$$

Note that the first derivative for y and second derivative for y would be the same with above for x by exchanging m_j and p_j.

One minor issue about the objective function in (5) is that the function is not concave. Thus, directly using Newton's method will suffer from numerical problem employed a pseudo-Newton's method [1] to overcome this non-concavity while maximally preserving the advantages of Newton's method. Instead of directly using the Hessian matrix H, we subtract a constant matrix to make it strictly negative definite, i.e., $H' = H - \delta I$ where I is an identity matrix. This modification to Newton's method enables the algorithm to converge smoothly to a local optimal for a non-concave problem.

Abbreviations

The abbreviations for Europe (**Fig. 2**) are: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia; Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; SZ, Switzerland; TR, Turkey; UA, Ukraine; YG, Yugoslavia.

The abbreviations for the globe map (**Fig. 3**) are: AG, Adygei; BL, Balochi; BK, BantuKenya; BS, BantuSouthAfrica; BQ, Basque; BD, Bedouin; BI, BiakaPygmy; BR, Brahui; BU, Burusho; CB, Cambodian; CL, Colombian; DA, Dai; DR, Daur; DZ, Druze; FR, French; HA, Han; HN,Han-NChina; HZ, Hazara; HE, Hezhen; IT, Italian; JP, Japanese; KL, Kalash; KT, Karitiana; LH, Lahu; MK, Makrani; MD, Mandenka; MY, Maya; MP, MbutiPygmy; ML, Melanesian; MI, Miao; MO, Mongola; MZ, Mozabite; NX, Naxi; OR, Orcadian; OQ, Oroqen; PL, Palestinian; PP, Papuan; PT, Pathan; PI, Pima; RU, Russian; SN, San; SD, Sardinian; SH, She; SD, Sindhi; SU, Surui; TU, Tu; TJ, Tujia; TC, Tuscan; UG, Uygur; XB, Xibo; YK, Yakut; YI, Yi; YR, Yoruba;

References

[1] Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, USA, (2000).