SUPPLEMENTARY MATERIAL

Selection and genotyping of unlinked genetic markers

A total of 37 unlinked SNPs, distributed across the entire genome and located outside any known gene regions, were selected for determining the contribution of European or African ancestry in the Cape Coloured population. The distance between adjacent markers on the same chromosome ranged from 100kb to 150Mb. Previously reported allele frequencies for the 37 SNPs were obtained from the SNPper database (www.snpper.chip.org) for African (Yoruban or African American) and European (CEPH or European American) populations [Supplementary Table 1] and were selected on the basis of allele frequency differences between these populations. The SNPs were genotyped using the matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry (Compact Sequenom MassARRAY^{TM}, Sequenom, San Diego, CA, U.S.A.) and the homogenous MassEXTEND chemistry. PCR and extension primer sequences were designed using Sequenom RealSNP (www.RealSNP.com). Multiplex (five to nine-plexes) PCR amplification was performed in a final volume of 5 ul for reactions containing 2.5 ng of DNA, 10X Qiagen HotStar Taq PCR buffer, 25 mM $MgCl_2$, 25 mM dNTPs, 200 nM of PCR primer (primer sequences and multiplex combinations are available upon request) and 0.15U Qiagen HotStar Taq Polymerase using universal PCR cycling conditions. The MassEXTEND reaction was performed using the appropriate termination mix, 600 nM of each extension primer (primer sequences available upon request) and 0.063U of Thermosequenase with cycling conditions as per Sequenom protocols.

f_2, f_3 and f_4 statistics

We have 4 distinct populations *W, X, Y, Z.* An allele has population frequencies *w, x, y, z* respectively We observe counts w_0,w_1 of the allele and the complementary allele in a sample from population *W*. Similarly we observe counts x_0,x_1; y_0,y_1; z_0,z_1. We will assume that the total count for each population is at least 2. Thus the natural (naive) estimator of *w* is

$$w' = \frac{w_0}{(w_0 + w_1)}$$

with similar definitions of x', y', z'. We wish to form unbiased estimates of quantities such as $(w - x)(y - z)$ which we term an f_4-*statistic*. It is easy to see that the naive estimate

$$f_4(W, X, Y, Z) = (w' - x')(y' - z')$$

Indeed is an unbiased estimator. Next suppose we want an estimator (f_3-*statistic*) for $(w - x)(w - y)$ where w appears twice. Consider the naive estimator:
$q = (w' - x')(w' - y')$. Then we can write q as,

$$q = ((w' - w) - (x' - x) + (w - x))((w' - w) - (y' - y) + (w - y))$$

This shows that the bias of q is $E(w' - w)^2$. Let $n_W = w_0 + w_1$ be the total allele count for W. Then

$$E(w' - w)^2 = \frac{w(1 - w)}{n_W}$$

Define $h_W = w(1 - w)$
($2\ h_W$ is the heterozygosity at the marker for population W). Then a natural estimator for h_W is

$$\hat{h}_W = \frac{w_0 w_1}{n_W(n_W - 1)} \qquad [1]$$

and we can readily check that $\hat{h}_W$ is unbiased. Putting this together we obtain:

$$f_3(W, X, Y) = (w' - x')(w' - y') - \hat{h}_W / n_W$$

and f_3 is an unbiased estimator of $(w - x)(w - y)$. Similarly we can define

$$f_2(W, X) = (w' - x')(w' - x') - \hat{h}_W/n_W - \hat{h}_X/n_X$$

and show that f_2 *(W, X)* is an unbiased estimator of $(w - x)^2$. In applications we always wish to compute weighted sums of the *f-statistics* across many markers. Unbiasedness is critical here ensuring convergence of our average *f-statistic* to the average we would obtain by using the true allele frequencies.

Scaling of our f_2, f_3 statistics

Our statistics resemble F_{st} with our f_2*-statistic* being essentially the numerator of the Cockerham-Weir estimator (1,2) of F_{st}. How we scale our statistics is irrelevant for our inference, but we prefer to use a fixed scaling so that f_2 becomes close to F_{st}. We computed F_{st} and f_2 (using Yoruba as an outgroup) for all pairs of populations in {*Coloured, Europe, SouthAsia, Bushmen, isiXhosa*} and then computed a scaling factor (population independent) s so as to minimize the square distance between F_{st} and sf_2. We obtain s = 0.293 and use this value in all calculations we report in this paper.

REFERENCES

1. Reynolds, J., Weir, B.S., Cockerham, C.C. (1983). Estimation of the coancestry coefficient: Basis for a short term genetic distance. *Genetics,* **105**, 776-779.
2. Weir, B.S., Cockerham, C.C. (1084). Estimating f-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.

Table S1. Screening of 37 unlinked SNPs within the Coloured (n = 268) and isiXhosa (n = 306) populations.

				Allele frequency			
	Chromosome			Reference[1]		Cape	
SNP	Band	Position	Allele	African	European	Coloured	isiXhosa
rs6679668	1p36.23	8090492	T	0.62	1.00	0.95	0.80
			C	0.38	-	0.05	0.20
rs753345	1q23.1	154741028	G	0.57	0.89	0.77	0.62
			A	0.43	0.11	0.23	0.38
rs300780	2p25.3	100819	G	0.50	0.49	0.52	0.50
			A	0.50	0.51	0.48	0.50
rs1213579	2p25.3	2001333	G	0.87	0.42	0.60	0.79
			A	0.13	0.58	0.40	0.21
rs1861497	2p25.1	8002736	A	0.89	0.31	0.64	0.90
			G	0.11	0.69	0.36	0.10
rs732892	2q14.2	119541979	C	0.40	0.90	0.75	0.53
			T	0.60	0.10	0.25	0.47
rs6442890	3p26.3	502223	A	0.40	0.52	0.54	0.56
			G	0.60	0.48	0.46	0.44
rs937803	3p24.1	30088664	C	0.98	0.66	0.88	0.99
			T	0.02	0.34	0.12	0.01
rs2968684	4p16.2	5007062	C	0.55	0.77	0.76	0.59
			T	0.45	0.23	0.24	0.41
rs7720419	5p15.33	642343	T	0.93	0.41	0.64	0.95
			A	0.07	0.59	0.36	0.05
rs7702150	5p15.33	1222112	G	0.48	0.84	0.68	0.65
			A	0.52	0.16	0.32	0.35
rs163587	5p13.2	35013904	C	0.64	1.00	0.77	0.51
			G	0.36	-	0.23	0.49
rs736864	6p25.3	131221	T	0.83	0.21	0.57	0.95
			G	0.17	0.79	0.43	0.05
rs1986345	6p25.3	730010	G	0.90	0.31	0.52	0.81
			C	0.10	0.69	0.48	0.19
rs399269	6p23	15005036	C	0.66	0.21	0.56	0.70
			T	0.34	0.79	0.44	0.30
rs2968858	7q36.1	150043936	A	0.68	0.52	0.52	0.46
			G	0.32	0.48	0.48	0.54
rs6558434	8p23.3	1201464	C	0.55	0.82	0.82	0.72
			A	0.45	0.18	0.18	0.28
rs6988580	8p23.2	5041500	G	1.00	0.65	0.84	0.99
			T	-	0.35	0.16	0.01
rs1548122	8q21.3	90009353	C	ND	0.51	0.52	0.54
			T	ND	0.49	0.48	0.46
rs1908233	9p24.3	549434	A	0.96	1.00	0.90	0.90
			G	0.04	-	0.10	0.10
rs4741213	9p24.3	1229776	G	0.88	0.48	0.67	0.96
			A	0.12	0.52	0.33	0.04
rs1328273	9p22.3	16013469	G	1.00	0.56	0.83	0.98
			A	-	0.44	0.17	0.02
rs1986466	9p21.1	30008156	T	0.76	0.48	0.57	0.66
			C	0.24	0.52	0.43	0.34
rs1598505	11p15.4	5007007	G	0.45	0.63	0.68	0.67
			C	0.55	0.37	0.32	0.33
rs923805	11p15.3	12008042	G	0.62	0.25	0.46	0.60
			A	0.38	0.75	0.54	0.40

rs868249	12p13.33	78147	T	0.47	0.88	0.68	0.51
			C	0.53	0.12	0.32	0.49
rs739973	12p13.33	1518835	G	0.53	ND	0.54	0.71
			A	0.47	ND	0.46	0.29
rs2532544	12p13.32	4004736	C	0.57	1.00	0.78	0.56
			T	0.43	-	0.22	0.44
rs1904239	12p13.2	12024132	G	0.67	ND	0.59	0.60
			A	0.33	ND	0.41	0.40
rs108990	16p13.3	1005434	T	0.43	0.82	0.60	0.46
			C	0.57	0.18	0.40	0.54
rs7193708	16p13.2	8024801	T	0.56	0.93	0.80	0.67
			C	0.44	0.07	0.20	0.33
rs1125988	16q12.1	50030013	C	0.57	0.50	0.59	0.42
			G	0.43	0.50	0.41	0.58
rs759974	17p13.3	347709	G	0.97	0.44	0.77	0.93
			A	0.03	0.56	0.23	0.07
rs1940658	18p11.32	2019280	C	0.82	0.43	0.61	0.82
			T	0.18	0.57	0.39	0.18
rs7244992	18p11.22	9004820	T	0.38	0.89	0.79	0.53
			C	0.62	0.11	0.21	0.47
rs7260021	19p13.2	9022607	G	0.28	0.61	0.55	0.45
			C	0.72	0.39	0.45	0.55
rs91710	19p13.11	18002123	A	0.62	0.44	0.69	0.74
			G	0.38	0.56	0.31	0.26

[1]Allele frequencies were obtained from the NCBI dbSNP database as available per May 2008 update. Allele frequencies for the African population are as reported for Yorubans from Ibadan, Nigeria (YRI), while allele frequencies for the European population are as reported for Caucasians from the United States with northern and western European ancestry (CEU). ND, Not determined.

Fig. S1. Analysis of 37 unlinked genetic markers for 306 isiXhosa (red), 268 Coloured (green) and 50 European (blue) South Africans. Outliers were excluded for genome-wide analysis.