Human Molecular Genetics, 2010, Vol. 19, No. 3 **411–419**
doi:10.1093/hmg/ddp505
Advance Access published on November 5, 2009

Genetic structure of a unique admixed population: implications for medical research

Nick Patterson[1,†], Desiree C. Petersen[2,†], Richard E. van der Ross[3], Herawati Sudoyo[4], Richard H. Glashoff[5], Sangkot Marzuki[4], David Reich[1,6] and Vanessa M. Hayes[2,7,*]

[1]Broad Institute of MIT (Massachusetts Institute of Technology) and Harvard University, Cambridge Center, Cambridge, MA, USA, [2]Cancer Genetics Group, Children's Cancer Institute Australia for Medical Research, Sydney Children's Hospital, Randwick, NSW 2031, Australia, [3]University of the Western Cape, Cape Town, South Africa, [4]Eijkman Institute for Molecular Biology, Jakarta, Indonesia, [5]Division of Medical Virology, Department of Pathology, Faculty of Health Sciences, University of Stellenbosch, Tygerberg, South Africa, [6]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA and [7]Faculty of Medicine, University of New South Wales, Randwick, NSW 2031, Australia

Received May 12, 2009; Revised September 29, 2009; Accepted October 29, 2009

Statement: In naming population groups, we think a chief aim is to use terms that the group members use themselves, or find familiar and comfortable. The terms used in this manuscript to describe populations are as historically correct as possible and are chosen so as not to offend any population group. Two of the authors (DCP and REvdR) belong to the Coloured population, with one of the authors (REvdR) having contributed extensively to current literature on the history of the Coloured people of South Africa and served as Vice-President of the South African Institute of Race Relations. According to the 2001 South African census (http://www.statssa.gov.za/census01/HTML/CInBrief/CIB2001.pdf), 'Statistics South Africa continues to classify people by population group, in order to monitor progress in moving away from the apartheid-based discrimination of the past. However, membership of a population group is now based on self-perception and self-classification, not on a legal definition. Five options were provided on the questionnaire, Black African, Coloured, Indian or Asian, White and Other. Responses in the category "Other" were very few and were therefore imputed'. We have elected to use the term Bushmen rather than San to refer to the hunter-gatherer people of Southern Africa. Although they have no collective name for themselves, this decision was based on the term Bushmen (or Bossiesman) being the more familiar to the communities themselves, while the term San is the more accepted academic classification.

Understanding human genetic structure has fundamental implications for understanding the evolution and impact of human diseases. In this study, we describe the complex genetic substructure of a unique and recently admixed population arising ∼350 years ago as a direct result of European settlement in South Africa. Analysis was performed using over 900 000 genome-wide single nucleotide polymorphisms in 20 unrelated ancestry-informative marker selected Coloured individuals and made comparisons with historically predicted founder populations. We show that there is substantial genetic contribution from at least four distinct population groups: Europeans, South Asians, Indonesians and a population genetically close to the isiXhosa sub-Saharan Bantu. This is in good accord with the historical record. We briefly examine the implications of determining the genetic diversity of this population, not only for furthering understanding of human evolution out of Africa, but also for genome-wide association studies using admixture mapping. In conclusion, we define the genetic structure of a uniquely admixed population that holds great potential to advance genetic-based medical research.

[*] To whom correspondence should be addressed. Tel: +61 293820229; Fax: +61 293821850; Email: vhayes@ccia.unsw.edu.au
[†] These authors equally contributed to this work.

INTRODUCTION

Many human populations are admixed, having been formed from mixing of two or more parental populations, which are genetically distinct. African-Americans are a familiar example, with on average ~20% of their genome from European ancestry (1). It is important to characterize the genetic make-up of admixed populations. In many cases this will shed light on human genetic history, although the admixture offers some technical opportunities for disease mapping (defined as the genome-wide average proportion of ancestry from a given population). In particular, the ancestry of local regions of the genome will often be very diverse in an admixed population, whereas cultural and environmental confounders may be relatively uniform. This concept of admixture mapping has, for example, been successfully utilized to map disease loci for type 2 diabetes in Hispanics (2), high blood pressure in African-Americans (3) and prostate cancer in African-American men (4).

A population with a complex history of admixture, is the Coloured population of South Africa. The people of South Africa, a nation of over 48 million, are broadly classified as 79.6% African, 9.1% European, 8.9% Coloured and 2.5% Indian/Asian (Statistics South Africa 2007, www.statssa.gov.za). The eleven official languages include nine African (isiNdebele, isiXhosa, isiZulu, Sepedi, Sesotho, Setswana, siSwati, Tshivenda and Xitsonga), English and Afrikaans (Dutch ancestrally linked language unique to South Africa). The vast majority of Coloured people (estimated 4.5 million) are historically Afrikaans speaking (although this is changing) and form the largest ethnic group (50.2%) residing in the Western Cape region (Statistics South Africa 2007, www.statssa.gov.za).

We need to discuss the historical events that resulted in the founding of the Coloured as a distinct genetic group. The Coloured emerged as a consequence of enslavement by European settlers at the Cape. The settlers were predominantly Dutch, with some French and German. Major ethnic contributions to the Cape slaves were made from East Africa (Mozambique), Madagascar (and surrounding islands), India (a variety of populations including Coromandel, Malabar, Bengal and Ceylon, now Sri Lanka) and substantial numbers from Indonesia (more specifically Java, Batak, Bali and Bugis and Makassar from the Celebes). Together with settler-slave unions, unions with the native Bushmanoid people where common. These included the Khoikhoi and Bushmen people (also known as San), commonly pooled together as the Khoisan due to their clicking languages. The Khoikhoi (meaning 'men of men' or 'real people') were taller, physically stronger, and were cattle owners, further distinguishing them from the hunter-gatherer Bushmen (5). Although both Khoikhoi and Bushmen have contributed to the Coloured gene-pool, the contribution of the Khoikhoi appears to have been more predominant (6,7). In 1808 with the abolishment of slavery, the term 'Coloured' was introduced and is still in use today. For more detailed information, refer to van der Ross (8).

In this study, we use genome-wide genetic analysis for almost 900 000 markers (Affymetrix SNP6.0 array) from the admixed (20 Coloured) and founder populations (20 isiXhosa and 20 Indonesian), together with HapMap data (9–11) and Human Genome Diversity Panel (HGDP) data (12,13) to examine the genetic contribution of the Coloured population. We describe a novel method related to linear regression for computing the degree of admixture given samples from an admixed population and samples from the populations believed to have contributed to the admixing. We find substantial genetic heterogeneity and address the implications of these findings not only to understand human evolution, but also medical research efforts, including (a) population stratification contribution to disease association studies, (b) the viability of admixture mapping for complex trait loci identification and (c) the impact of population diversity in advancing drug development strategies.

Table 1. Pairwise F_{ST} (X 1000) between the populations in this study

	COL	XHO	SAN	YRI	CEU	CHB	SAS	IND
COL		62	127	72	40	75	26	65
XHO			77	18	155	188	136	177
SAN				100	213	245	194	236
YRI					152	183	132	174
CEU						110	20	102
CHB							79	20
SAS								72
IND								

RESULTS

Genetic diversity between study populations

Table 1 shows F_{ST} estimates for our sample population (20 Coloureds, COL) and various hypothesized populations (Fig. 1) contributing to Coloured admixture (20 isiXhosa, XHO; 20 Indonesian, IND; 4 Southern African Bushmen, SAN; 46 South Asians, SAS; 58 Utah Europeans, CEU; 55 Yoruba West Africans, YRI; 44 Han Chinese, CHB). Note that F_{ST} between the isiXhosa and HapMap Yoruba ($F_{ST} = 0.018$) is substantially smaller than between the isiXhosa and Bushmen ($F_{ST} = 0.077$). The Coloured are closer to the South Asian sample ($F_{ST} = 0.026$) than to any other sample population. We failed to detect within-Indonesia variation in our samples that is relevant to the genetics of the Coloured. Our Indonesian samples do show significant though small differences (Table 2), with the Java sample clearly different from the other populations, and the Batak Toba and Makassar also being differentiated from each other.

Genetic ancestry of the coloured people

The substantial scatter of our Coloured samples across both axes of the principal components plot (Fig. 2A), shows that the Coloured have diversity both on an African–non-African axis and on an axis only related to non-African genetics. On a plot where axes were formed from the HapMap Yoruba and Bushmen populations, we demonstrate that both the South African Coloured and isiXhosa populations have admixture and population diversity on a Bantu–Bushmen axis (Fig. 2B). The within Yoruba scatter on this plot is not statistically significant. As the admixture is similar, this

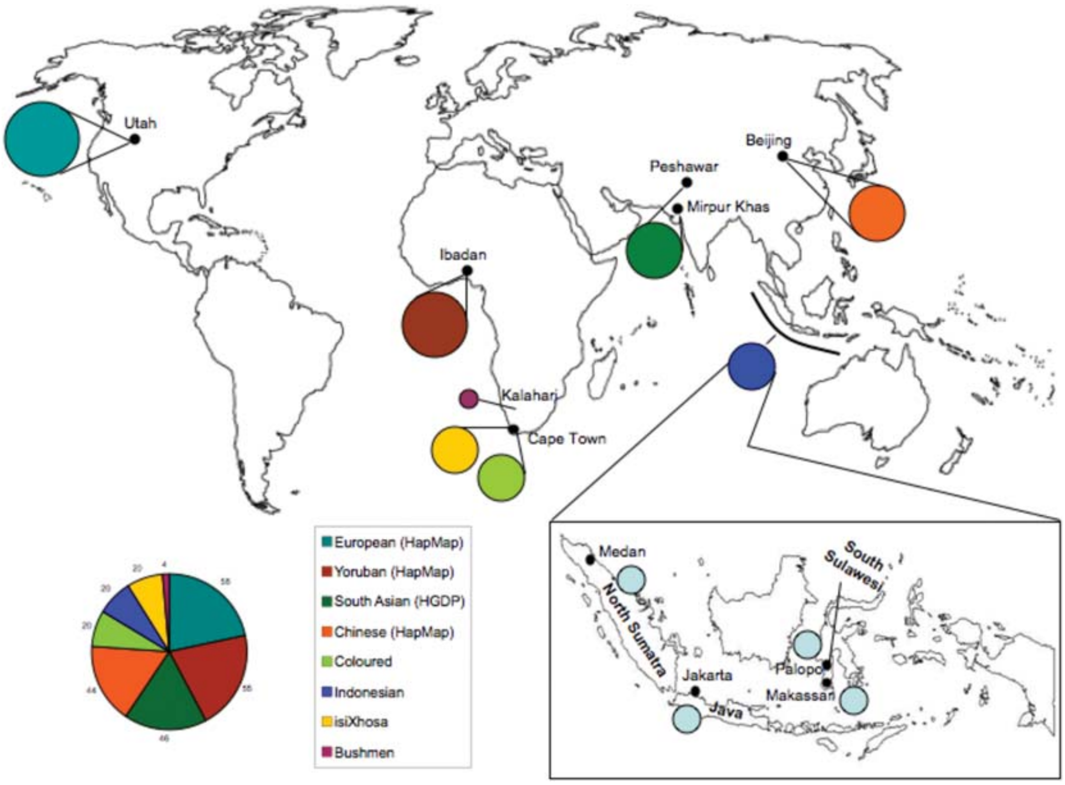

Figure 1. Map of the world depicting population contributions. The European, Yoruba and Han Chinese sample collections were from the HapMap project, the South Asian population from the Human Genome Diversity Project (HGDP) and the remaining populations were collected as part of this study. The Coloured and isiXhosa were from the same geographical location in South Africa, although the Indonesian population was made-up of Batak-Toba (North Sumatra), Javanese, Bugis (South Sulawesi) and Makassar.

Table 2. Pairwise F_{ST} (X 1000) between the Indonesian populations in this study

	Barak-Toba	Bugi	Java	Makassar
Barak-Toba		2	5	4
Bugi			6	0
Java				6
Makassar				

suggests that the African component of the Coloured population is closely related to the isiXhosa. Using the European and Chinese HapMap data, together with the South Asian populations to form the axes, we show that the Coloured population has some European ancestry, ancestry related to Indonesia or China, and possibly some South Asian ancestry (Fig. 2C). When forming our axes using our Indonesian samples and the HapMap Chinese, our Coloured samples are notably shifted towards the Indonesian end of the first component (Fig. 2D). This demonstrates a genetic relatedness of the Coloured to Indonesia, in accord with historical evidence. However, within-Indonesia divergence is not related in a detectable way to our samples, either because the Indonesians taken to South Africa by the Dutch were rather homogeneous, or because subsequent gene-flow within the Coloured has tended to reduce the diversity. We found no evidence of any Chinese or other East Asian ancestry in our Coloured samples. Plotting our data against the South Asian and HapMap European data sets, we show the Coloured to be shifted towards the European population relative to all other populations (Fig. 2E). We regard this as decisive proof of a European component to the admixture. Hence we have demonstrated very strong evidence based on the genetic data of admixture from a European, and Indonesian and an African (isiXhosa related) population.

Interpretation of genetic diversity data within the South Asian populations tends to be more complicated as a result of ancestral West Eurasian (close genetically to modern Europe) and other Asian admixture. We therefore chose to carry out a principal components analysis of European (Hapmap CEU), isiXhosa, Indonesian and South Asian populations, and then project the Coloured samples on this axis (Fig. 3). This plot confirms a highly variable African ancestry in the Coloured samples. We find the plot suggestive of a South Asian component of ancestry in the Coloured, but this is not entirely decisive and we confirm the South Asian component using our regression based methods which are more quantitative and lead to precise statistical tests.

Determining the mixing coefficients

Applying our regression-style technique we can estimate the mixing coefficients in our sample. We take our source (mixing) populations to be isiXhosa, European, South Asian

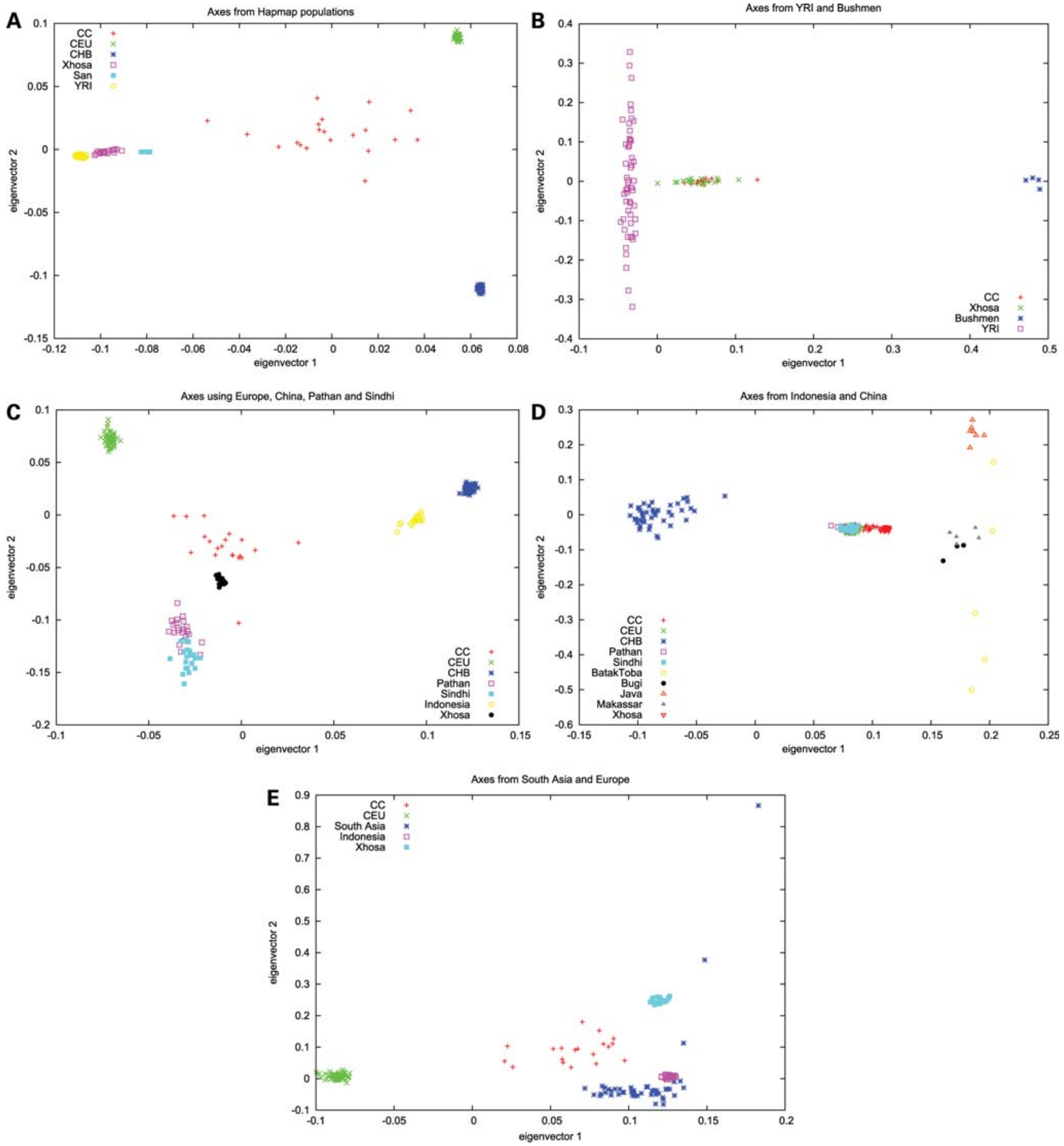

Figure 2. Principal components scatter plots. Indicating (**A**) our Coloured and isiXhosa samples across an African versus non-African axis, (**B**) a Bantu versus Bushmen axis, (**C**) European versus Asian (Han Chinese, Indonesian and South Asian) axis, (**D**) Chinese versus Indonesian axis and (**E**) European versus South Asian axis.

and Indonesian. We get estimated mixing coefficients when applying our method to the Coloured, separately on the autosomes and X-chromosome (Table 3). To further clarify the meaning of these admixture coefficients, an example can be taken from our observed estimated mixing proportions of Indonesians, on our autosomal data, to be 0.180. We interpret this as meaning that if we pick a random autosomal locus, and a random chromosome from our Coloured sample (not from all Coloured in South Africa) and follow the line of descent back 40 generations (much earlier than the admixing events), then the probability that the line would be in a chromosome of an individual from Indonesia is ∼18%.

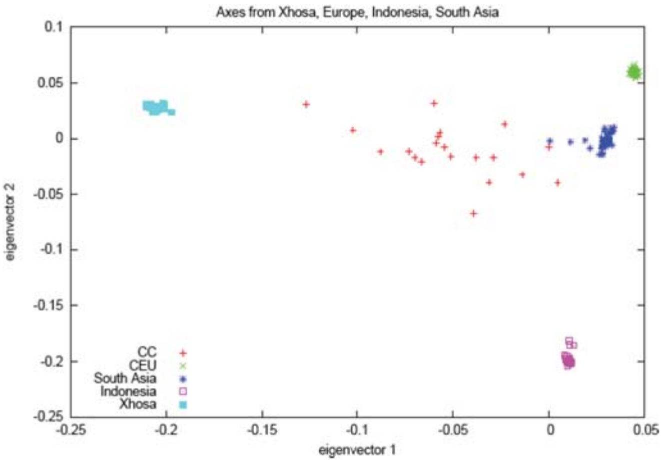

Figure 3. Principal components scatter plot. Indicating our 20 Coloured samples projected onto an axis of European (Hapmap CEU), isiXhosa, Indonesian and South Asian populations.

Table 3. Mixing coefficients on autosomes and X-chromosome

Population	Autosome	Std. error	X-chromosome	Std. error
European	0.231	0.008	0.140	0.036
South Asian	0.221	0.009	0.240	0.049
Indonesian	0.180	0.004	0.234	0.017
isiXhosa	0.369	0.003	0.387	0.013

We obtained an error covariance for the autosomes from which the standard errors were taken (Table 4), using a weighted block jackknife (14,15). This allows computation of a Z-score (mean 0 variance 1 if there is no real predictive improvement). Note that these are standard errors for mixing coefficients of our Coloured sample. The Coloured population is genetically so variable that speaking of the distribution of ancestry coefficients is hardly meaningful without a very precise description of the population being sampled. The error covariance is not close to diagonal. This is expected since the coefficients are constrained to sum to one. Notably there is a strongly negative cross-term between South Asia and Europe (-61.403×10^{-6}). These two populations are genetically quite close ($F_{ST} = 0.026$) and so our 'regression' finds the sum of the coefficients of South Asia and Europe are much better determined than the difference.

We note that there are highly significant differences between our mixing coefficients on the X-chromosome compared with the autosomes. The reason for this must be gender-biased gene flow in our mixing population. We observe greatly reduced European contribution to X-chromosome markers, and a strongly increased Indonesian contribution, probably caused by mating between European males and Indonesian females, in agreement with historical records.

We wanted to confirm that our Coloured samples contain genetic admixture from South Asia. We divided the genome into 5 cM blocks and chose half the blocks at random. We built 'optimal' models with and without South Asia, and estimated the error of the best model on the remaining blocks. This gave square error estimates of 0.0034 and 0.0040, respectively. To test significance, we computed mean and standard error of the difference of the mean-square estimated prediction error, block by block and then computed the standard error of the difference using a weighted jackknife. This showed that incorporating South Asia yields a Z-score of 9.46 against a null model with South Asia excluded. As a negative control we tried using China (CHB) instead of our South Asian samples. The coefficient for CHB was 0.003 with a standard error of 0.008 and the Z-score, using the prediction errors, was 0.098.

Table 4. Covariance of mixing coefficients (× 1 000 000)

	European	South Asian	Indonesian	isiXhosa
European	**57.993**	−61.403	4.931	−1.521
South Asian		**86.607**	−18.041	−7.163
Indonesian			**16.331**	−3.221
isiXhosa				**11.905**

Bold, standard errors.

DISCUSSION

It is important to understand the genetic origins of admixed populations. In many societies, the proportion of individuals with significant recent genetic admixture is on the increase due to increased urbanization and migration (16). Increased population admixture influences genome heterozygosity, which in turn will affect phenotypes relevant to health. Thus genetic admixture has many implications in medical research (as reviewed in 17). Using high-density genome-wide genotyping, we have described an admixed population from South Africa, the Coloured, with substantial genetic heterogeneity [in contrast to the findings of Barreiro *et al.* (18)] and show at least approximately the proportions of population ancestry of each of our samples.

We want to expand on what we mean by 'population' in our study. It is obvious that we can only be discussing historical gene-flow in the samples that we examine. The population labels are self-identified and certainly our genetic analysis shows clearly that the labels are meaningful in that the labelling indeed distinguishes our groups from one another. Our principal components technique is blind to the labelling, in the sense that the coordinates we plot in our figures are produced without using the labels, and the labels are used simply as annotation to aid interpretation. More precisely, we often do use a subset of the samples to choose axes, and then plot the samples chosen and additional samples on these axes. Thus in some sense each principal components picture has two kinds of samples, those used to make the axes and the rest. We are clear about which samples are in which category. In many of our 'populations' the samples are quite homogeneous (for example HapMap Yoruba or Europeans) and it is reasonable to think of them as random samples from a much larger set. However, two of our sample groups, the isiXhosa and Coloured, show large genetic variation which we interpret as different proportions of ancestors from ancestral population groups as the genealogy of individuals show varying proportions of recently admixing populations. Here the meaning of 'population' is less clear. To further discuss the admixture of the samples we are analyzing, and the extent to which this is typical of the genetics of other individuals with the same self-identification, would require much further analysis and data collection.

The concept of gender-biased gene flow described in this study is a common pattern observed in human history. Examples are reported in Yemen (19), Latinos from Columbia (20), Icelanders (21) and African-Americans (22). In our case we observe greatly reduced European contribution to the X-chromosome and a strongly increased Indonesian contribution. This suggests mating between European males and Indonesian females. The initial European settlers to arrive at the Cape were predominantly men and unions were common between male settlers and female slaves or natives. As slaves were prohibited to marry (a law only abolished in 1823) the men either bought the slave their freedom (named a Free Black) or they bore an unwed child. From 1652 to British occupation in 1796, a total of 1350 settler-Free Black marriages were recorded whereas only 137 Free Black marriages were recorded. Marriage was freely possible with the non-slaved Bushmanoid populations. Almost all recorded mixed marriages were between a settler male and either a Free Black or Khoisan female (8).

Our results show that as expected from the historical evidence, the Coloured population has a complex genetic history. There is four-way admixture from a population related to isiXhosa, Europeans, South Asians and Indonesians. As the isiXhosa are themselves also admixed, showing evidence varying across individuals of relationship to Bushmen, though the largest genetic component in the isiXhosa is Bantu, we can say that the Coloured are five-way admixed (at least). Our study is in fair agreement with recently published microsatellite and in/del data (23). This complex genetic structure is a critical consideration for medical research strategies.

Like many other recognized population groups of the world, the Coloured people have been used for determining genetic-based disease association. Our results imply that studies of association to disease in this population need to carefully consider population stratification so as to avoid false disease-risk allele associations (24). Only a single group have addressed population stratification in a medical study of the Coloured, where a genetic-based tuberculosis association study reported no significant population stratification for 25 unlinked markers (18). This is in contrast to our findings which clearly show large and important stratification. We therefore call for applying adjustments for population stratification when using this population to determine genetic-based disease associations. One suitable technique is EIGENSTRAT (25).

Admixed populations can be used to identify loci that contribute to ethnic variation in complex disease risk, using admixture mapping (26–28). In comparison with standard genetic-based association studies admixture mapping requires fewer markers. This powerful tool also has a greater statistical power than family-based linkage studies (29). One of our major objectives of this study was to determine the viability of the Coloured population for this gene mapping approach. Our results show that the Coloured admixture is so complex that techniques of admixture mapping where a small number of markers are used, informative for ancestry, will be hard to carry out for technical reasons. However, the availability of high content arrays will allow these technical challenges to be addressed, ultimately presenting the uniquely admixed Coloured as a valuable population for medical genetic studies. Many risk alleles will be at very different frequencies in different human populations, either because of random genetic drift or because the populations have been under different selection pressures, perhaps due to differing environments. This is evident in Supplementary Material, Table S1 where for example a number of alleles have no differences in frequency between the Coloured and say the European population, although others demonstrate a significantly altered frequency. In a population such as the Coloured the complex genetics may well prove an advantage, as a risk allele may be at high frequency in one of the admixing populations, making it easy to find compared with a scan in a homogeneous population where it is at low frequency. It's worth pointing out here that although admixture mapping will be difficult, a whole genome association study is technically straightforward, even in a population with substantial stratification, by using methods previously described (25). Thus this population may provide a unique opportunity to study ethnic based differences to complex diseases.

In addition to the direct implications to medical research, understanding population genetic diversity patterns is important to understanding the evolution of modern humans and thus in turn the evolution of human disease. To truly understand human evolution it is important not only to understand genetic diversity of numerous populations, but also to investigate these diversities on a genome-wide level. A recent multi-locus nested clade analysis has shown how limiting single gene or DNA regions, including the unisexually inherited mitochondrial and Y-chromosome regions, are in capturing human evolutionary history (30). The 'out-of-Africa' hypothesis places sub-Saharan African populations at the source of this origin with outward flow to the rest of Africa (13).

In conclusion, using a novel method for computing degree of admixture, we demonstrate clear evidence of African (genetically close to isiXhosa), Indonesian, European and South Asian contributions in our recently (~350 years) admixed Coloured samples. This agrees well within the historical evidence. Therefore the Coloured people represent a new class of unique genomes created from a divergent genetic background, including more than one of the described six major ancestral human genetic clusters (31). This admixture holds strong potential to offer new insights into complex gene–gene and gene–environment interactions, insight into human evolution and human disease evolution, and enabling medical research efforts unparalleled by any other population.

MATERIALS AND METHODS

Ethics statement

Although the data were analyzed anonymously (only ethnic background and gender taken into consideration), informed consent was obtained from the South African study participants and the Ethics Review Committee of the University of Stellenbosch, South Africa (98/158), as well as the University of New South Wales, Australia Human Research Ethics Committee (HREC #08244), approved the study protocol.

Study populations and sample selection

The Southern African populations in this study include the Coloured, isiXhosa and Bushmen. The Coloured (COL) and isiXhosa (XHO), sampled from the Western Cape region of South Africa (Fig. 1), were recruited as part of a control arm for a larger study investigating population-specific genetic variation for comparisons in disease association. All were blood donors from the Western Province Blood Transfusion Service. The isiXhosa population is from the Nguni-speaking nation of Africans and after the Coloured, is the largest population group in the Western Cape. The four Bushmen (SAN) included in this study represent the indigenous population of South Africa at the dawn of the Coloured people.

The 20 Coloured and 20 isiXhosa samples assessed in this study were selected from a ‘pre-analysis’ of 37 selected unlinked ancestry informative markers distributed across the genome (Supplementary Material, Table S1). This pre-analysis was critical for sample study selection as self-attempts at altering population classification, thus social status and opportunity, during the South African era of race segregation (apartheid) was common practice. Using this data we could exclude population outliers, and also identify probable relatedness among our samples (Supplementary Material, Fig. S1, $n = 268$ Coloured, 306 isiXhosa and 50 European South Africans). This pre-selection allowed for an increased study power based on a within population sample size of 20. Further selection criteria included restriction to female participants. Although our sample size for the Bushmen is small, they are useful for inference. With whole-genome data, even small samples are highly informative for many population genetic applications, as the very large number of loci means that we are in fact sampling a substantial number of independent ancestral genomic regions.

The Indonesian (IND) contribution to the Coloured population was assessed by genotyping 20 randomly selected samples from three regions including five Batak-Toba of north Sumatra, three Bugis and five Makassar of south Sulawesi (formerly known as Celebes) and seven Javanese of central Java (Fig. 1). The four populations included in this study, from the more than 17 000 islands that made up Indonesia and 483 ethnic groups, were selected based on historical accounts of Indonesian contribution to slavery or political exiles to the Cape during Dutch settlement. These samples were collected with informed consent as part of a larger study to examine the population structure of the Indonesian archipelago, reviewed and approved by the Eijkman Institute Research Ethics Committee. The genetic relationships between the Indonesian and other ethnic populations of Southeast Asia have been established through a 50 000 genome-wide single nucleotide polymorphisms (SNPs) study to map Human Genetic history in Asia (HUGO Pan-Asian SNP Consortium, unpublished data).

Historical information led us to believe that the Pathan and Sindhi populations were likely to be genetically close to the South Asian (SAS) ancestors of the Coloured people and therefore 46 samples from the HGDP (12,13) were also included (Fig. 1). Using available International HapMap Project data (8–10) we were able to analyze 58 European (CEU, Utah residents with ancestry from northern and western Europe), 55 West African (YRI, Yoruba from Ibadan, Nigeria) and 44 Chinese (CHB, Han Chinese from Beijing) (Fig. 1).

Whole-genome genotyping

Genotyping was performed using the Affymetrix Genome-Wide Human SNP Array 6.0, containing 906 600 SNPs. These SNPs are derived from previous Affymetrix arrays and include tag SNPs from the International HapMap project. SNPs were analyzed for 20 Coloured, 20 isiXhosa and 20 Indonesian samples with a call accuracy rate of 99.71%. In brief, using the Affymetrix Genome-Wide Human SNP Nsp/Sty Assay Kit 5.0/6.0, a minimum of 500 ng of genomic DNA (at a concentration of 50 ng/μl), resuspended in nuclease-free reduced EDTA TE buffer, was digested with *Nsp*I and *Sty*I restriction enzymes followed by ligation to adaptors that recognize the cohesive 4 base pair (bp) overhangs. The different sized fragments, resulting from the restriction enzyme digestion, serve as the substrates for adaptor ligation. A generic primer recognizing the adaptor sequence is used to amplify adaptor-ligated DNA fragments ranging from 200 to 1100 bp in length. The amplified products for each restriction digest are combined and purified using polystyrene beads. Following fragmentation and labelling of the amplified DNA, it is hybridized to a Genome-Wide Human SNP Array 6.0 in Affymetrix GeneChip Hybridization Oven 640. After washing and staining in the Affymetrix GeneChip Fluidics Station 450, the arrays were scanned with the Affymetrix GeneChip Scanner 3000 7G (full protocol is supplied by the manufacturer, Affymetrix, Santa Clara, CA, USA).

Statistical analysis

Although the sample size for some of our populations are small, whole-genome genotyping is highly informative for many population genetic applications. Here we are genotyping nearly 900 000 SNPs and obtain tight estimates of quantities such as F_{ST}. We are sampling here a substantial number of independent ancestral genomic regions. Indeed, as we discuss briefly below, statistics such as F_{ST} can be computed meaningfully from a single sample.

We use two main techniques. The first uses a selected set of populations to compute principal components axes (32). We then project other populations on the axes computed. This is a powerful method for determining population relationships using principal components plots (Figs 2 and 3). The title indicates which populations were used to compute the axes, with the remaining populations plotted by projection onto the axes. Our second method is novel, though related to linear regression. We wish to compute the degree of admixture given samples from an admixed population, and samples from the populations believed to be admixing.

The problem we wish to solve is that we have populations $\{Mi\}$ $(i = 1, 2,\ldots N)$ and a variant allele with population frequency x_i for population i. Given another population A in which the allele has frequency x we wish to estimate x as:

$$x \approx \sum_{i=1}^{N} w_i x_i$$

where $\sum_{i=1}^{N} w_i = 1$. Of course if A is an admixed population with admixing populations M_i this would yield an estimator for the mixing coefficients. If $x_i^{[k]}$, $i = 1, \ldots, N$ and $x^{[k]}$ are known for many alleles $k = 1,\ldots K$ then it would be natural to solve a regression problem: Minimize

$$S = \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\left(w_i x_i^{[k]} - x^{[k]}\right)^2}{V_k} \qquad (1)$$

where V_k is some estimate of an error variance and we require $\sum_i w_i = 1$. A reasonable procedure would be to choose an outgroup population, obtain an empirical estimator P of the allele frequency in the outgroup and set $V_k = P\ (1 - P)$. We can introduce Lagrange multipliers and check that at a minimum of S in Eq. (1) we have:

$$\sum_j C_{ij} w_j = \lambda$$

where

$$C_{ij} = \left(\frac{1}{K}\right) \sum_{k=1}^{K} \frac{\left(x^{[k]} - x_i^{[k]}\right)\left(x^{[k]} - x_i^{[k]}\right)}{V_k}$$

for some λ. Thus we minimize S by solving the linear equations

$$\sum_j C_{ij} u_j = 1$$

and then setting

$$w_j = \frac{u_j}{\sum_j u_j}. \qquad (2)$$

Our problem is more complicated, because we are not given $x_i^{[k]}$ and $x^{[k]}$. Instead we have allele counts $a_i^{[k]}$ and $b_i^{[k]}$ of the variant and reference alleles. Similarly we have corresponding counts $a^{[k]}$ and $b^{[k]}$ for the variant and reference alleles in the admixed population. Our idea is that we will estimate the ideal coefficients C_{ij}. We show in supplementary material that we can give unbiased estimators $E_{ij}^{[k]}$ for $\left(x^{[k]} - x_i^{[k]}\right)\left(x^{[k]} - x_j^{[k]}\right)$. In the notation of our appendix:

$$E_{ij}^{[k]} = f_3(A, M_i, M_j)$$

at allele k if $(i \neq j)$ otherwise:

$$E_{ii}^{[k]} = f_2(A, M_i).$$

Indeed, using the Lehmann–Scheffé theorem (Theorem 4.2.2) (33) $E_{ij}^{[k]}$ is a uniformly minimum variance unbiased estimator. We choose V_k throughout this paper as $V_k = P(1 - P)$ where P is the empirical allele frequency for the (HapMap) Yoruba. Then it follows that

$$C'_{ij} = \frac{1}{K} \sum_k E_{ij}^{[k]}$$

will be very close to C_{ij} for K large. We choose w_i by solving

$$\sum_j C'_{ij} u_j = \lambda$$

and then setting w_i as in Eq. (2), thus substituting C' for C in our algorithm for determining w_j for the case of known allele frequencies.

The problem we are solving is quite different from the problem of estimating the best predictor of x given allele counts from other populations. Here we are trying to compute the optimal mixing coefficients that we would obtain in an idealized situation where the counts were very (infinitely) large so that the population frequencies are known exactly. Our idea is especially appropriate for situations where the sample sizes are small, so that the naive estimator for $x_i^{[k]}$

$$\frac{a_i^{[k]}}{a_i^{[k]} + b_i^{[k]}}$$

will have large standard error, but we have a very large number of markers. Our f_2 and f_3 statistics resemble the numerator of the familiar F_ST estimators [see for instance Weir and Cockerham (34)]. How we scale these statistics is not important for inference, but we choose a linear scaling that makes our f_2 statistics close to our F_ST statistics.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank Steve McCarroll for permission to use Affymetrix 6.0 genotyping data from the haplotype map samples, and Stephan C. Schuster, Pennsyvania State University, for useful communications. We acknowledge the participants in this study. We thank Sr Heather Money and staff for the co-ordination and collection of blood specimens from the Western Province Blood Transfusion Service, South Africa.

Conflict of Interest statement. None declared.

FUNDING

Additional infrastructure support for this work was provided by the Australian Cancer Research Foundation (ACRF) and the Ramaciotti Centre at the University of New South Wales, Australia. The supplementary pre-analysis data was performed while D.C.P. and V.M.H. were located at the Garvan Institute of Medical Research, Sydney, Australia. V.M.H. is a Cancer Institute of New South Wales Fellow, Australia. This project was supported by funding from the Medical Research Council (MRC) of South Africa, Poliomyelitis Research Foundation of South Africa, and Hyperion Asset Management, Australia.

REFERENCES

1. Destro-Bisol, G., Maviglia, R., Caglià, A., Boschi, I., Spedini, G., Pascali, V., Clark, A. and Tishkoff, S. (1999) Estimating European admixture in African Americans by using microsatellites and a microsatellite haplotype (CD4/Alu). *Hum. Genet.*, **104**, 149–157.
2. Parra, E.J., Hoggart, C.J., Bonilla, C., Dios, S., Norris, J.M., Marshall, J.A., Hamman, R.F., Ferrell, R.E., McKeigue, P.M. and Shriver, M.D. (2004) Relation of type 2 diabetes to individual admixture and candidate gene polymorphisms in the Hispanic American population of San Luis Valley, Colorado. *J. Med. Genet.*, **41**, e116.
3. Zhu, X., Luke, A., Cooper, R.S., Quertermous, T., Hanis, C., Mosley, T., Gu, C.C., Tang, H., Rao, D.C., Risch, N. and Weder, A. (2005) Admixture mapping for hypertension loci with genome-scan markers. *Nat. Genet.*, **37**, 177–181.
4. Freedman, M.L., Haiman, C.A., Patterson, N., McDonald, G.J., Tandon, A., Waliszewska, A., Penney, K., Steen, R.G., Ardlie, K., John, E.M. *et al.* (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. USA*, **103**, 14068–14073.
5. Boonzaaier, E., Malherbe, C., Smith, A. and Berens, P. (1996) *The Cape Herders: A History of the Khoikhoi of Southern Africa*. Ohio University Press, Ohio, USA.
6. Morris, A.G. (1997) The Griqua and the Khoikhoi: biology, ethnicity and the construction of identity. *Kronos J. Cape History*, **24**, 106–118.
7. Balson, S. (2007) *Children of the Mist, the lost tribe of South Africa*. Interactive Presentations Pty Ltd, Australia.
8. van der Ross, R.E. (2005) *Up from Slavery: Slaves at the Cape, their origins, treatment and contribution*. Ampersand Press and University of the Western Cape, Cape Town.
9. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1229–1320.
10. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
11. The International HapMap Constortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–862.
12. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R. *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.
13. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. and Myers, R.M. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.
14. Künsch, H.R. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, **17**, 1217–1241.
15. Busing, F.M.T.A., Meijer, E. and van der Leeden, R. (1999) Delete-*m* jackknife for unequal *m*. *Stat. Comput.*, **9**, 3–8.
16. Rudan, I., Carothers, A.D., Polasek, O., Hayward, C., Vitart, V., Biloglav, Z., Kolcic, I., Zgaga, L., Ivankovic, D., Vorko-Jovic, A. *et al.* (2008) Quantifying the increase in average human heterozygosity due to urbanisation. *Eur. J. Hum. Genet.*, **16**, 1097–1102.
17. Copper, R.S., Tay, B. and Zhu, X. (2008) Genome-wide association studies: implications for multiethnic samples. *Hum. Mol. Genet.*, **17**, R151–R155.
18. Barreiro, L.B., Neyrolles, O., Babb, C.L., Tailleux, L., Quach, H., McElreavey, K., Helden, P.D., Hoal, E.G., Gicquel, B. and Quintana-Murci, L. (2006) Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis. *PLoS Med.*, **3**, e20.
19. Richards, M., Rengo, C., Cruciani, F., Gratrix, F., Wilson, J.F., Scozzari, R., Macaulay, V. and Torroni, A. (2003) Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. *Am. J. Hum. Genet.*, **72**, 1058–1064.
20. Bedoya, G., Montoya, P., García, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., Hedrick, P.W. and Ruiz-Linares, A. (2006) Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. *Proc. Natl. Acad. Sci. USA*, **103**, 7234–7239.
21. Helgason, A., Sigurethardottir, S., Gulcher, J.R., Ward, R. and Stefansson, K. (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am. J. Hum. Genet.*, **66**, 999–1016.
22. Lind, J.M., Hutcheson-Dilks, H.B., Williams, S.M., Moore, J.H., Essex, M., Ruiz-Pesini, E., Wallace, D.C., Tishkoff, S.A., O'Brien, S.J. and Smith, M.W. (2007) Elevated male European and female African contributions to the genomes of African American individuals. *Hum. Genet.*, **120**, 713–722.
23. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O. *et al.* (2009) The genetic structure and history of Africans and African Americans. *Science*, **324**, 1035–1044.
24. Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N. *et al.* (2004) Assessing the impact of population stratification on genetic association studies. *Nat. Genet.*, **36**, 388–393.
25. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
26. Hoggart, C., Shriver, M., Kittles, R., Clayton, D. and McKeigue, P. (2004) Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.*, **74**, 965–978.
27. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D. *et al.* (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.*, **74**, 979–1000.
28. Darvasi, A. and Shifman, S. (2005) The beauty of admixture. *Nat. Genet.*, **37**, 118–119.
29. McKeigue, P.M. (2005) Prospects for admixture mapping of complex traits. *Am. J. Hum. Genet.*, **76**, 1–7.
30. Templeton, A.R. (2007) Genetics and recent human evolution. *Evolution*, **61**, 1507–1519.
31. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.
32. Patterson, N., Price, A. and Reich, D. (2006) Population Structure and Eigenanalysis. *PLoS Genet.*, **2**, e190.
33. Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco, USA.
34. Weir, B.S. and Cockerham, C.C. (1984) Estimating f-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.