Heredity (2007) 98, 151–156

www.nature.com/hdy

ORIGINAL ARTICLE

Is urbanization scrambling the genetic structure of human populations? A case study

M Ashrafian-Bonab, LJ Lawson Handley and F Balloux
Theoretical and Molecular Population Genetics Group, Department of Genetics, University of Cambridge, Cambridge, UK

Recent population expansion and increased migration linked to urbanization are assumed to be eroding the genetic structure of human populations. We investigated change in population structure over three generations by analysing both demographic and mitochondrial DNA (mtDNA) data from a random sample of 2351 men from 22 Iranian populations. Potential changes in genetic diversity (θ) and genetic distance (F_{ST}) over the last three generations were analysed by assigning mtDNA sequences to populations based on the individual's place of birth or that of their mother or grandmother. Despite the fact that several areas included cities of over one million inhabitants, we detected no change in genetic diversity, and only a small decrease in population structure, except in the capital city (Tehran), which was characterized by massive immigration, increased θ and a large decrease in F_{ST} over time. Our results suggest that recent erosion of human population structure might not be as important as previously thought, except in some large conurbations, and this clearly has important implications for future sampling strategies.
Heredity (2007) **98,** 151–156. doi:10.1038/sj.hdy.6800918; published online 15 November 2006

Keywords: population structure; human populations; mtDNA; Iran; demography

Introduction

The human population has grown dramatically over the last 100 years and there has been a corresponding increase in migration to urban areas owing to better healthcare, education and work opportunities. It has been suggested that demographic expansion and migration are scrambling the genetic structure of human populations (Cavalli-Sforza *et al*., 1991) and as a response, projects have been initiated to sample the genetic diversity of the world's populations before their genetic identity is lost for good (Cann *et al*., 2002, The Human Genome Diversity Project, HGDP, and more recently the Genographic Project). It remains unclear though whether these recent phenomena have already produced changes in the genetic structure of human populations and whether the suggestion that human populations are homogenizing into a 'global melting pot' is truly justified.

This is an important question from an intrinsic or historical perspective (e.g. for understanding the complexity of human migration and demographic history) but is also relevant to applied research in human genetics and medicine. In targeted drug administration, association studies and studies of gene function for example, even small amounts of admixture can lead to false-positive results and failure to detect genuine associations (e.g. Pritchard and Rosenberg, 1999; Deng *et al*., 2002; Hirschhorn *et al*., 2002; Freedman *et al*., 2004; Marchini *et al*., 2004; Helgason *et al*., 2005).

Although obtaining ancestral information from sampled individuals is quite a widespread practice (particularly for association studies e.g. Ardlie *et al*., 2002), this information has rarely been used to investigate changes in population structure over time (see Helgason *et al*., 2005 for an exception). Knowing the place of birth of a sampled individual, their parents and grandparents, provides a means to test whether increased migration during recent decades really has eroded human population structure. An important related question is whether increased migration (and the assumed erosion of population structure) is a general phenomenon or whether it is confined to major conurbations such as capital cities. Immigration to major cities, which seems a fair expectation, will have important consequences for sampling because limitations owing to cost and accessibility mean that samples are often obtained from major cities and assumed to be representative.

Here, we explore the hypothesis that in recent generations, population expansion and increased migration (linked to urbanization) have eroded signals of human population structure. We performed a case study of the Iranian population, using an intensive, unbiased sampling procedure and collecting information on the birthplace of sampled individuals, parents and grandparents. In the last three generations, the Iranian population has more than tripled in size to its present tally of >68 million. Iran is a large country (>1.6 million km^2), and its populations, which are highly diverse in terms of culture and language, were relatively isolated geographically until the second half of twentieth century. Since then, major highways and railroads have been

Correspondence: Dr LJ Lawson Handley, Theoretical and Molecular Population Genetics Group, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK.
E-mail: ljl27@cam.ac.uk
Received 23 June 2006; revised 7 September 2006; accepted 16 September 2006; published online 15 November 2006

constructed to connect population centres, and the proportion of the population living in urban areas has more than doubled (from 27% in 1950 to 64% in 2000, United Nations Population Division). Migration, particularly from rural areas towards major cities, is therefore expected to have increased substantially in recent generations.

Methods

We sampled 2351 unrelated men from 22 populations, which comprehensively represent the diversity of Iran in terms of culture, language and geography (Figure 1, Table 1). Samples were randomly collected from consenting volunteers at transfusion centers to ensure an unbiased sampling strategy. Information was obtained on the place of birth of the sampled individual (generation t) and on the place of birth of their parents (generation t_{-1}) and grandparents (generation t_{-2}). As mitochondrial DNA (mtDNA) does not recombine, we could simply assign sequences to populations in each generation using place of birth of the sampled individual, their mother or grandmother. Such a strategy is equivalent to random sampling in the two previous generations as long as the sample does not comprise a large proportion of siblings and cousins, a situation we have been careful to avoid.

DNA was extracted using the PAXgene blood DNA kit (Preanalytix, GmbH, Germany). MtDNA control region sequences (both HVSI and HVSII) were obtained following polymerase chain reaction and sequencing with standard primers and protocols (Torroni *et al*., 1996; Mogentale-Profizi *et al*., 2001).

Genetic diversity was computed within populations (theta from the number of segregating sites, $\theta_s = 2Ne\mu$, where Ne is the effective population size and μ the mutation rate, Watterson, 1975) and genetic structure estimated among populations (population specific F_{ST}; Wright, 1969) using DnaSP v4 (Rozas *et al*., 2003) and ARLEQUIN v3 (Excoffier *et al*., 2005) respectively. Both θ_s and F_{ST} were obtained per population for each of the three generations and then averaged over populations. The mtDNA sequence data generated in this article are freely available from the authors on request.

Results

Figure 2 illustrates the percentage of randomly sampled individuals from each locality whose grandparents and/or parents were born in the same area, compared to those individuals that were born in the sampling locality, but whose parents or grandparents were born outside the area ('immigrants'). For simplicity, if the sampled individual and their grandparents were born in the same locality, it was assumed that the parents were also born in that locality (this was the case for all but three individuals in the whole sample set). Patterns of immigration in the last three generations are similar for female and male individuals (Figure 2a and b, respectively), although there is a very slight tendency for male subjects to migrate more than female subjects (over all populations $\chi^2 = 4.77$, 1 d.f., $P < 0.05$). Immigration has been low in the majority of populations (less than 10% for all but three populations). In striking contrast, the majority of individuals sampled in Tehran (FA1) have parents or grandparents from outside the area (62% have immigrant mothers or grandmothers and 69% fathers or grandfathers). In the majority of cases, immigrants come from adjacent populations, and have therefore traveled very short distances (less than 300 km, data not shown).

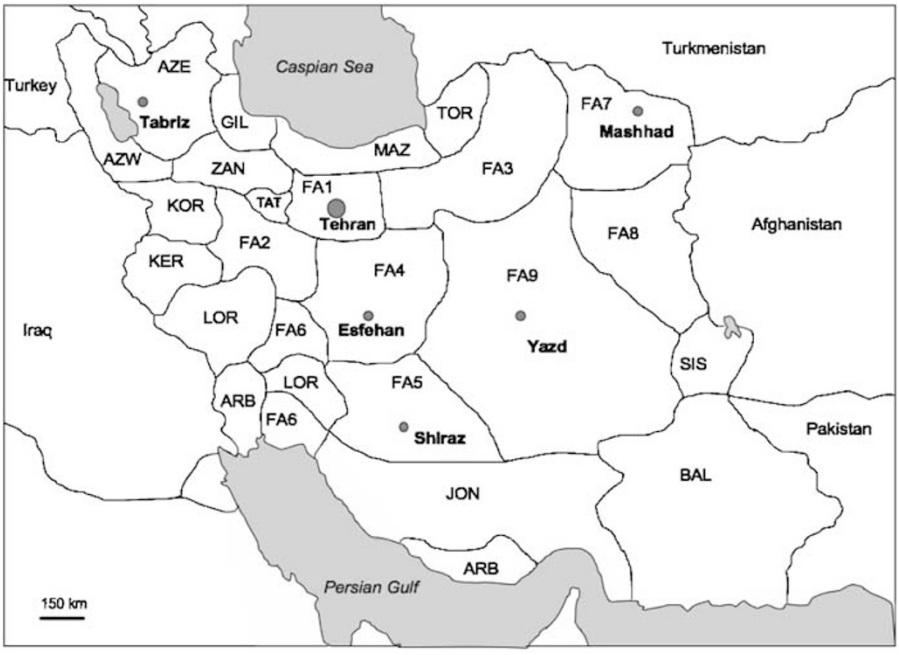

Figure 1 Sampling information. Population codes and sample sizes (n = number of individual samples collected in each locality) are as in Table 1 ('KOR': Kords from Kordestan, not included in present study). The large dot represents the capital, Tehran (>10 million inhabitants), smaller dots are large conurbations ⩾one million inhabitants. Borders in the figure roughly correspond to cultural regions based on ethnicity and language (see Ethnologue homepage for more information).

Table 1 Sampling information

Population code	*Population name*	*Locality*	*Sample size*
ARB	Arab	Khuzestan and Boushehr Provinces	96
AZE	Azeri	East Azerbaijan and Ardabil Provinces	110
AZW	Azeri	West Azerbaijan Province	102
BAL	Balochi	Sistan and Balochestan Province (Balochi area)	135
FA1	Fars	Tehran and Qazvin Provinces	390
FA2	Fars	Qom, Markazi and Hamedan Provinces	75
FA3	Fars	Semnan Province	55
FA4	Fars	Esfehan Province	105
FA5	Fars	Shiraz Province	45
FA6	Fars	North Khorasan Province	70
FA7	Fars	South Khorasan Province	50
FA8	Fars	Khuzestan Province	52
FA9	Fars	Yazd and Kerman Provinces	91
GIL	Gilaki	Gilan Province	125
JON	Jonobi	Hormozghan Province	100
KER	Kerman	Kermanshah Province	120
LOR	Lorestani	Lorestan Province	125
MAZ	Mazandrani	Mazandaran Province	120
SIS	Sistani	Sistan and Balochestan Province (Sistani area)	145
TAT	Takestan	Qazvin Province	70
TOR	Torkaman	Golestan Province	90
ZAN	Zanjani	Zanjan Province	80
Total			2351

Although immigrants to Tehran have traveled further, immigration declines with distance and effectively ceases after 1000 km. In Tehran there is less than a 50% chance of sampling an mtDNA sequence that was present two generations ago, and approximately a one in six chance of sampling an Azeri East mtDNA (data not shown), which illustrates that caution is needed when sampling in the capital city.

Mean within-population genetic diversity is very similar over the three generations ($\bar{\theta}_{s(t-2)} = 0.369$, $\bar{\theta}_{s(t-1)} = 0.369$, $\bar{\theta}_{s(t)} = 0.361$, one-way repeated measures analysis of variance (ANOVA) $F = 1.926$, d.f. $= 2$, $P = 0.178$, not significant, Figure 3a). However three populations do show deviation from this pattern. A slight increase in θ_s over time is observed in the Balochi population (BAL $\theta_{s(t-2)} = 0.027$, $\theta_{s(t-1)} = 0.029$ and $\theta_{s(t)} = 0.035$) and in the area of Masshad (FA6 $\theta_{s(t-2)} = 0.030$, $\theta_{s(t-1)} = 0.032$ and $\theta_{s(t)} = 0.036$). By contrast a marked increase in genetic diversity over time is seen in Tehran (FA1 $\theta_{s(t-2)} = 0.043$, $\theta_{s(t-1)} = 0.047$, $\theta_{s(t)} = 0.057$ and Figure 3a).

There is a slight decrease in mean pairwise population specific $\bar{F}_{ST}$ per generation (Figure 3b: $\bar{F}_{ST(t-2)} = 0.0140$, $\bar{F}_{ST(t-1)} = 0.0133$, $\bar{F}_{ST(t)} = 0.0130$ when Tehran is excluded from the calculations). Although this change is small, the effect is highly significant (one-way repeated measures ANOVA $F = 35.77$, d.f. $= 2$, $P < 0.001$). Note that this effect is not detected if F_{ST} is computed for each population against all other populations pooled together. A quantitatively more substantial decrease in F_{ST} over time is seen in Tehran (Figure 3b: Tehran (FA1) $F_{ST(t-2)} = 0.0128$, $F_{ST(t-1)} = 0.0120$ and $F_{ST(t)} = 0.010$).

Discussion

During the twentieth century the average annual population growth rate in Iran was among the highest in the world (peaking at ~3.6% between 1976 and 1986 in relation to the slackening of family planning laws following the Islamic Revolution in 1979). The increase in population size, as well as development of transport, industry, education and health services, is expected to have had a significant impact on recent migrations. Long distance migrations to major cities are assumed to have increased but more local scale migrations are also expected. We therefore predicted that increased migration would be apparent in our demographic data and translate to increased genetic diversity within populations and lowering of genetic distance among populations over time (i.e. decreased population structure). In contrast with this prediction we found no difference in within-population diversity over time for most populations, and a quantitatively small, but significant decrease in population structure (Figure 3). Only the capital, Tehran, showed considerable demographic and genetic evidence of immigration (Figures 2 and 3). In fact, sampling at random in Tehran would produce only a one in two chance of selecting an individual with native parents and grandparents. This result has important consequences for other sampling strategies, which often concentrate on capitals or other large cities, and reiterates the need to avoid major cities as sources of samples for population genetic analyses (e.g. Bowcock and Cavalli-Sforza, 1991).

Patterns of migration between men and women generally suggest greater patrilocality and female-biased dispersal on local and regional scales (e.g. Salem *et al*., 1996; Seielstad *et al*., 1998; Mesa *et al*., 2000; Thomas *et al*., 2000; Oota *et al*., 2001; Wilson *et al*., 2001; Wen *et al*., 2004). By contrast, the demographic data presented here indicates that, whereas patterns of male and female individual migration to Tehran and overall are similar (Figure 2) there is a slight tendency for men to migrate more often and further than women (over all populations $\chi^2 = 4.77$, 1 d.f., $P < 0.05$). That this difference in migration patterns is only slight, suggests that our main mtDNA results should also hold for Y chromosomes.

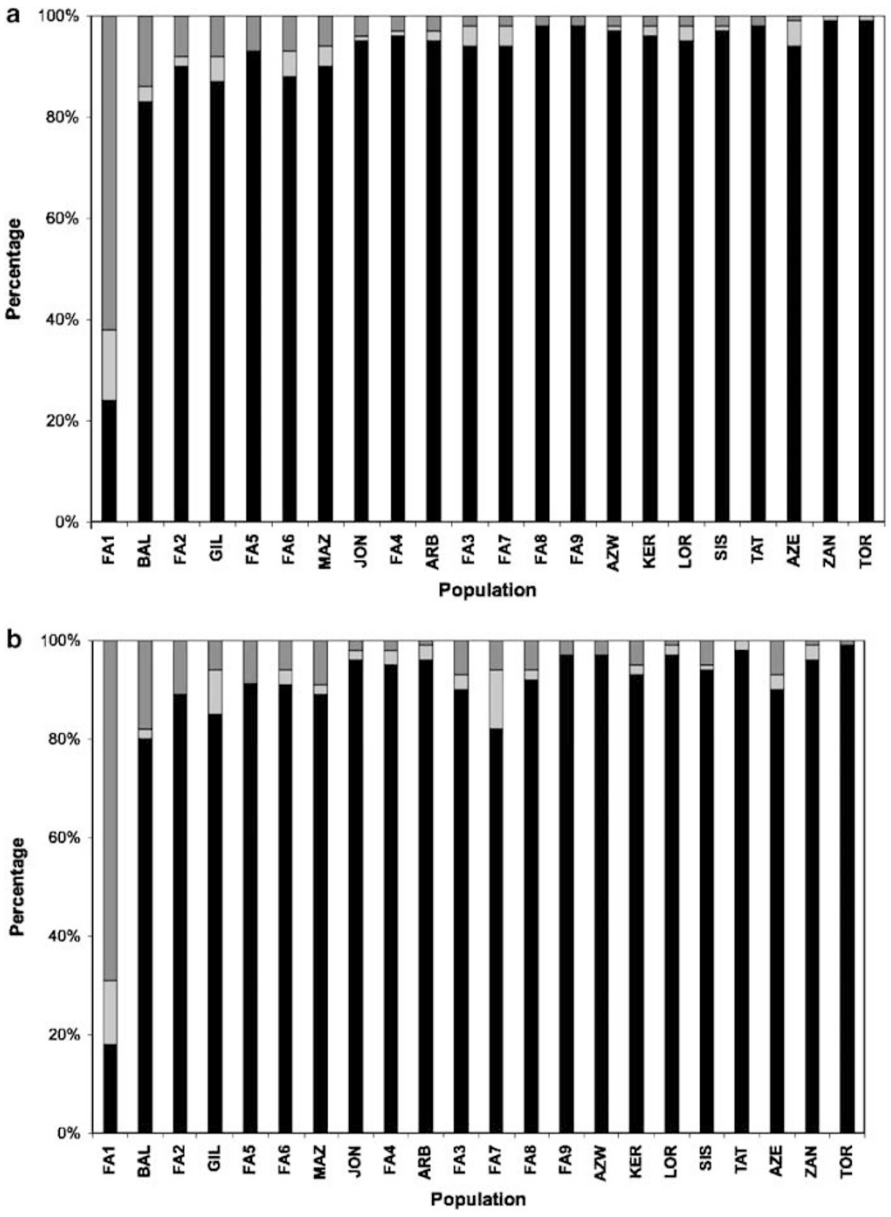

Figure 2 Demographic data. Percentage of (**a**) female and (**b**) male immigrants per population. Dark grey: individuals with immigrant parents, light grey: individuals with immigrant grandparents, black: individuals with both parents and grandparents born locally. Population order corresponds to the decreasing percentage of individuals with immigrant mothers (**a**). Order of populations in (**b**) is the same as for (**a**) to allow direct comparison.

Gathering data on an individual's place of birth clearly provides a unique insight into human population structure over time, and we highlight the importance of collecting this type of information during sampling. We expected recent migration within Iran to be considerable owing to rapid population growth and increased urbanization. However, despite extensive sampling, and inclusion of several major conurbations (six cities with over one million inhabitants), both our genetic and demographic data are consistent with only slight erosion of population structure in the last three generations, except in the capital city. These results do not therefore support the idea that human population structure has been extensively scrambled in recent decades, but do illustrate that sampling in capital cities (and preferably other very large cities) should be avoided unless ancestral information is obtained.

The dramatic population expansion and increased urbanization seen in Iran during the twentieth century, as well as the size and location of the country, make it a particularly suitable case study for investigating recent changes in human population structure. The demographic processes that have shaped the diversity of the Iranian population are similar to those that have shaped the whole of the Middle East region. Among-population variation in this region is low compared to the world-wide average (1.3% compared to 5.4%, Rosenberg *et al*., 2002), which suggests that populations are subject to more migration (or have a recent common ancestry, which seems unlikely) than in more remote parts of the

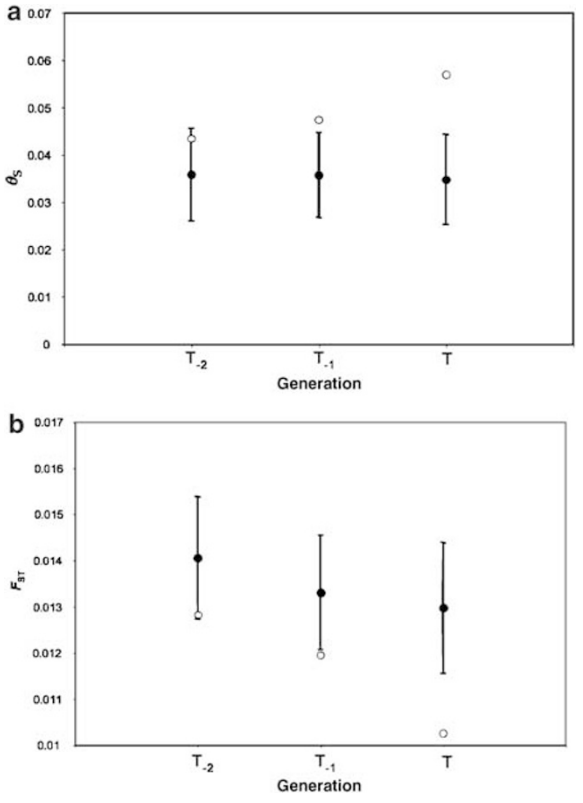

Figure 3 Genetic data: (**a**) Mean within-population genetic diversity (θ_s) per generation and (**b**) mean population specific $\bar{F}_{ST}$ per generation calculated from mtDNA. Black circles: mean ($\pm$s.d. error bars) for all populations except FA1 (Tehran); grey circles: FA1.

world. Finally, the rate of urbanization in Iran is similar to that of other developing countries (e.g. China, Brazil, Philippines, Argentina, United Nations Population Division). Therefore, while the implications of this case study may not be appropriate for the human population as a whole, we expect them to be applicable on a broad scale, to other developing countries with similar demographic histories.

Acknowledgements

We thank Laurent Lehmann, Hua Liu, Andrea Manica, Towfique Raj, Steve Russell, Camille Szmaragd, Chris Tyler-Smith, for discussion and suggestions, two anonymous reviewers for their helpful comments and the BBSRC of Great Britain for funding.

Web resources

Ethnologue http://www.ethnologue.com/

Genographic Project, https://www3.nationalgeographic.com/genographic/

Human Genome Diversity Project, http://www.stanford.edu/group/morrinst/hgdp.html

United Nations Populations Division, http://www.un.org/esa/population/unpop.htm

References

Ardlie KG, Lunetta KL, Seielstad M (2002). Testing for population subdivision and association in four case–control studies. *Am J Hum Genet* **71**: 304–311.

Bowcock A, Cavalli-Sforza L (1991). The study of variation in the human genome. *Genomics* **11**: 491–498.

Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L *et al.* (2002). A Human Genome Diversity Cell Line Panel. *Science* **296**: 261b–262b.

Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Deegan RM, King M-C (1991). Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics* **11**: 490–491.

Deng H-W, Gao G, Li J-L (2002). Estimation of deleterious genomic mutation parameters in natural populations by accounting for variable mutation effects across loci. *Genetics* **162**: 1487–1500.

Excoffier L, Laval G, Schneider S (2005). Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evolut Bioinformatics Online* **1**: 47–50.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N *et al.* (2004). Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**: 388–393.

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005). An Icelandic example of the impact of population structure on association studies. *Nat Genet* **37**: 90–95.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002). A comprehensive review of genetic association studies. *Genet Med* **4**: 45–61.

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004). The effects of human population structure on large genetic association studies. *Nat Genet* **36**: 512–517.

Mesa NR, Mondragon MC, Soto ID, Parra MV, Duque C, Ortiz-Barrientos D *et al.* (2000). Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre- and post-Columbian patterns of gene flow in South America. *Am J Hum Genet* **67**: 1277–1286.

Mogentale-Profizi N, Chollet L, Stevanovitch A, Dubut V, Poggi C, Pradie MP *et al.* (2001). Mitochondrial DNA sequence diversity in two groups of Italian Veneto speakers from Veneto. *Ann Hum Genet* **65**: 153–166.

Oota H, Settheethamishida W, Tiwawech D, Ishida T, Stoneking M (2001). Human mtDNA and Y-chromosomal variation is correlated with matrilocal versus patrilocal residence. *Nature* **29**: 20–21.

Pritchard JK, Rosenberg NA (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**: 220–228.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA *et al.* (2002). Genetic structure of human populations. *Science* **298**: 2381–2385.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Salem AH, Badr FM, Gaballah MF, Pääbo S (1996). The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. *Am J Hum Genet* **59**: 741–743.

Seielstad MT, Minch E, Cavalli-Sforza LL (1998). Genetic evidence for a higher female migration rate in humans. *Nat Genet* **20**: 278–280.

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M *et al.* (2000). Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba – the 'black Jews of Southern Africa'. *Am J Hum Genet* **66**: 674–686.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R *et al.* (1996). Classification of European mtDNAs

from an analysis three European populations. *Genetics* **144**: 1835–1850.
Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 256–276.
Wen B, Li H, Lu DR, Song XF, Zhang F, He YG *et al*. (2004). Genetic evidence supports demic diffusion of Han culture. *Nature* **431**: 302–305.
Wilson JF, Weiss DA, Richards DA, Thomas MG, Bradman N *et al*. (2001). Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proc Natl Acad Sci USA* **98**: 5978–5983.
Wright S (1969). *Evolution and the Genetics of Populations II*. The Theory of Gene Frequencies University of Chicago Press: Chicago.