Evidence that two main bottleneck events shaped modern human genetic diversity

W. Amos* and J. I. Hoffman

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

There is a strong consensus that modern humans originated in Africa and moved out to colonize the world approximately 50 000 years ago. During the process of expansion, variability was lost, creating a linear gradient of decreasing diversity with increasing distance from Africa. However, the exact way in which this loss occurred remains somewhat unclear: did it involve one, a few or a continuous series of population bottlenecks? We addressed this by analysing a large published dataset of 783 microsatellite loci genotyped in 53 worldwide populations, using the program 'BOTTLENECK'. Immediately following a sharp population decline, rare alleles are lost faster than heterozygosity, creating a transient excess of heterozygosity relative to allele number, a feature that is used by BOTTLENECK to infer historical events. We find evidence of two primary events, one 'out of Africa' and one placed around the Bering Strait, where an ancient land bridge allowed passage into the Americas. These findings agree well with the regions of the world where the largest founder events might have been expected, but contrast with the apparently smooth gradient of variability that is revealed when current heterozygosity is plotted against distance from Africa.

Keywords: bottleneck; out of Africa; microsatellite; demography; heterozygosity; human

1. INTRODUCTION

Many studies have been aimed at understanding the origin and distribution of human genetic diversity (Cavalli-Sforza *et al.* 1994; Harpending & Rogers 2000; Ramachandran *et al.* 2005). Arguably the most striking and consistent pattern reported to date is a decline in neutral variability with increasing distance from Africa, found not only in genetic markers (Harpending & Rogers 2000; Prugnolle *et al.* 2005) but also in morphological traits (Manica *et al.* 2007; Betti *et al.* 2008) and even commensal bacteria (Linz *et al.* 2007). This decline reflects loss of diversity in population bottleneck(s) as we expanded out of Africa to colonize the world, but the timing, number and location of these events remain obscure.

The question of how many bottlenecks account for the distribution of modern human diversity has been relatively little studied (Rogers & Harpending 1992) and yields conflicting results. First, simulations indicate that the observed pattern is consistent with a linear stepping-stone model featuring a long series of founder events (Ramachandran *et al.* 2005; Liu *et al.* 2006). However, this does not preclude equally good fits based on other models. Equally, at the other extreme, large steps in single nucleotide polymorphism diversity between adjacent populations have been used to argue for two dominant bottlenecks, one 'out of Africa' and one around the Bering land bridge where humans crossed into the Americas (Hellenthal *et al.* 2008). The latter event is supported by both mitochondrial data (Wallace *et al.* 1985; Fagundes *et al.* 2008) and data from a few

nuclear markers (Hey 2005). However, mitochondrial sequences only inform on female lineages, while the adjacent population approach is least reliable in regions like the Bering Strait where population samples are extremely sparse. With limited support for two such contrasting models, further work is desirable.

Microsatellites can provide an alternative view on population bottlenecks. These comprise short tandemly repeated arrays of 2–6 bp motifs that evolve mainly by gaining and losing single repeat units (Weber & Wong 1993; Xu *et al.* 2000; Huang *et al.* 2002), the so-called stepwise mutation model (SMM; Kimura & Ohta 1978). Much theory has been developed around the SMM, leading to an understanding of the expected allele length distribution (Di Rienzo *et al.* 1994) and its relationship with population size changes (Kimmel *et al.* 1998). Based on this, Luikart *et al.* (1998) developed a method to infer historical bottlenecks using current genetic samples based on the following rationale. During a bottleneck, rare alleles tend to be lost quickly, before heterozygosity has been eroded. The result is a transient imbalance between heterozygosity and allele number (Luikart *et al.* 1998). To detect this imbalance, stochastic simulations are used to derive the expected relationship between heterozygosity and allele number at equilibrium, allowing any given marker to be assessed for whether it exhibits excess heterozygosity given its allele number, an approach implemented by the program BOTTLENECK.

One problem with the BOTTLENECK test is that microsatellites do not follow a strict SMM. Known deviations include mutation biases favouring expansion or contraction (Xu *et al.* 2000), interruption mutations within the repeat tract that slow the rate of slippage (Jin *et al.* 1996; Kruglyak *et al.* 1998), occasional larger 'jump' mutations of several repeat units (Di Rienzo *et al.* 1994; Schlötterer *et al.* 1998) and some form of

* Author for correspondence (w.amos@zoo.cam.ac.uk).

upper length boundary that prevents indefinite expansion (Amos & Clarke 2008). However, in terms of the BOTTLE-NECK test, the key variable, regardless of how it comes about, is the frequency of homoplasic mutations. The BOTTLENECK program therefore allows the user to specify a range of mutation models from a strict SMM, where homoplasy is frequent, through varying proportions of jump mutations, to the alternative extreme, the infinite alleles model (IAM) where every mutation is novel.

Here, we apply the program BOTTLENECK to a large published dataset of 783 microsatellite markers geno-typed in the Human Genome Diversity Cell Line Panel, comprising 53 populations worldwide (Ramachandran *et al.* 2005; Rosenberg *et al.* 2005). By using such a large panel, we hope to quantify in each population the relative strength of signal of a bottleneck, allowing fine-scale mapping of how colonization of the world eroded our neutral genetic variability.

2. MATERIAL AND METHODS

Data for 783 microsatellites genotyped in 53 worldwide populations were downloaded from http://rosenberglab.bioinformatics.med.umich.edu/diversity.html (Rosenberg *et al.* 2003, 2005). Bottleneck analysis was conducted using BOTTLENECK v. 1.2.02 (http://www1.montpellier.inra.fr/URLB/bottleneck/pub.html) (Luikart *et al.* 1998). We explored a range of mutation models, from the strict SMM through three models with varying proportions of jump mutations (two-phase models, TPMs), to the IAM. For the TPM we used the default variance of 30 and three different proportions of jump mutations: 10, 5 and 2 per cent.

The program BOTTLENECK tests each locus for an excess of heterozygosity per allele relative to a population at mutation–drift equilibrium. It does this by dividing the difference between the observed and expected heterozygosity by the standard error of the null distribution. Both the expected heterozygosity and the standard error are obtained by simulation using one of a range of possible mutation models. To assess which factors are most associated with a strong signal of a bottleneck, we constructed a general linear model (GLM) in 'R' v. 2.0.1 (http://www.r-project.org/). Our assumption is that, while the absolute significance of any given *t*-value remains unknown because it depends on knowing which mutation model is correct, the GLM will test for significant variation in mean *t* across populations, thereby identifying geographical regions where the signal is stronger or weaker than elsewhere. Predictor variables were: (i) distance from Africa and distance from Africa squared to capture non-linear geographical patterns; (ii) motif type, coded as a factor with four levels (the three commonest repeat motifs, AC, ATT, GATA plus all others), included in case the different motifs evolve differently; (iii) heterozygosity in Africa, taken as the Biaka Pygmy population, to reflect the mutation rate of the marker; and (iv) log modern population size to reflect modern demography. The minimum adequate model was established by fitting all terms plus all second-order interactions, and then sequentially removing terms that did not cause a significant reduction in deviance explained.

The raw '*t*-values' produced by BOTTLENECK are negatively skewed, ranging from less than two down to increasingly negative values as the size of the heterozygosity excess increases. To reduce the skew and to generate a scale on which large values indicate stronger evidence of a bottleneck, the raw *t*-values were transformed by taking $\log(2-t)$, where *t* is the raw *t*-value generated by BOTTLENECK. The transformed values yield normally distributed residuals when fitted as the response in a GLM. Thus, while the interpretation of any given *t*-value depends on both measurement error and the mutation model used to generate the null distribution, variation in mean *t*-value across diverse populations should provide a good measure of the relative strength of evidence for a bottleneck.

3. RESULTS

All loci in all populations were tested using all five mutation models. To explore the impact of using different mutation models, we calculated the Pearson correlation coefficient between untransformed *t*-values from the SMM against the corresponding *t*-values derived for each of the other four models. All correlations were strong, with r^2 values ranging from 0.857 to 0.992, indicating that rank order significance was essentially conserved across all mutation models. Thus, while the *t*-values are greatest for the IAM and smallest for the SMM, in terms of the relative magnitude of evidence for a bottleneck, the choice of model matters little. Consequently, the results we present are based on a model used widely by others, the TPM with 98 per cent single-step changes.

To investigate which factors most influence the magnitude of the transformed *t*-values, we fitted a GLM. Monomorphic locus–population combinations return a null value and were excluded. After model simplification, the minimum adequate model retained all predictor variables and is summarized in table 1. Despite explaining only 1.91 per cent of the null deviance, the large sample size means many of the terms are highly significant. Using this model we extracted fitted values for how transformed *t* varies with distance from Africa and heterozygosity (figure 1*a*). At low heterozygosity, more or less a single peak is present at the African end of the graph. As heterozygosity increases, the pattern shifts strongly towards one featuring a single peak located around 19 000 km from Africa, approximating the location of the Bering land bridge. To visualize the confidence intervals around the shape of the fitted surface, we used two extreme values of heterozygosity, 0.3 and 0.9, and for each we constructed an *XY* plot of the fitted values against distance from Africa with standard error obtained from the model (figure 1*b* and figure 1*c*, respectively). These errors indicate that the overall form of the fitted surface is robustly defined.

Because the fitted model contains only linear and quadratic terms, it cannot uncover multimodal or otherwise complicated patterns. For a more detailed look at the data, we used local regression to fit a smoothed spline to the fitted response, implemented using the 'locfit' package in R (figure 2). As above, a more informative view including 95 per cent confidence intervals was obtained by classifying the data into a series of heterozygosity bins and then fitting smoothed splines to data from each bin (figure 3). A complicated series of peaks and

Table 1. General linear model of evidence of a bottleneck across 53 worldwide populations genotyped for 783 microsatellite markers. (Explanatory variables fitted in the full model were distance from Africa and distance from Africa squared, motif type, heterozygosity in Africa, log modern population size and all second-order interactions. Only significant terms remaining in the reduced models are shown. The χ^2 values for each term represent the change in deviance after removing that term and all interactions involving that term from the model. Degrees of freedom (d.f.) associated with deletion of the term from the model.)

term	estimate	χ^2	d.f.	p
distance from Africa	-9.27×10^{-5}	13.11	7	<0.0001
(distance from Africa)^2	3.15×10^{-9}	10.53	7	<0.0001
population size	-0.0091	12.68	6	<0.0001
motif	—	38.45	15	<0.0001
heterozygosity	-0.75	16.24	6	<0.0001
distance from Africa : (distance from Africa)^2	-3.10×10^{-14}	15.23	1	<0.0001
distance from Africa : population size	3.23×10^{-6}	8.89	1	0.0029
distance from Africa : motif	—	9.38	3	<0.0001
distance from Africa : heterozygosity	8.11×10^{-5}	26.71	1	<0.0001
(distance from Africa)^2 : population size	-9.37×10^{-11}	7.90	1	0.0049
(distance from Africa)^2 : motif	—	8.60	3	<0.0001
(distance from Africa)^2 : heterozygosity	-1.88×10^{-9}	12.38	1	0.0004
population size : motif	—	4.10	3	0.0065
motif : heterozygosity	—	11.74	3	<0.0001

troughs is revealed, summarized as follows. At low heterozygosity there is a dominant peak in Africa, a trough in East Europe (7500 km) and a peak in eastern East Asia (EEA; approx. 12 500 km). At medium heterozygosity, where most markers lie, there is an out of Africa peak, a central Asian trough (CA; approx. 9000 km) and a broad double peak spanning EEA to central America (20 000 km), the EEA peak being more pronounced. In the highest heterozygosity classes, the out-of-Africa peak shifts towards Europe, the broad EEA–CA peaks remain and the trough shifts towards East Asia (11 000 km).

The strong African peak seen at the lowest heterozygosity might reflect an ascertainment bias: markers with low heterozygosity in Africa yet more variability elsewhere probably include some or many that have lost variability in Africa, whether through an African bottleneck or perhaps through chance linkage to gene(s) under selection, and hence these are likely to give a strong bottleneck signal. Using similar arguments it is difficult to see any heterozygosity measure that is entirely uncoloured, but for an alternative view we repeated the analyses using mean heterozygosity across all loci (see the electronic supplementary material, figure S1). Here, generally similar peaks and troughs are seen, but the strong African peak is now replaced at low heterozygosity by a strong peak seen in Europe. As before, this might be an artefact. Low worldwide heterozygosity will tend to indicate a marker that is unusual in Europe, because marker development tended to select for loci with high variability.

The above analyses all assume that the different parts of the world that are equidistant from our origin are equivalent and that is clearly not the case. For example, populations in the Middle East are nearer to the origin than some of the within-Africa populations. To get a view on how the individual populations behave, we plotted each major geographical group separately, with the mean $\pm$ 1 s.e. *t*-value for each separate population (figure 4). A clear pattern is seen, with Africans showing least evidence of a bottleneck and the Middle East on average the strongest evidence. Putting the Africans aside, there is a strong decline in mean t from the

Middle East through Europe and central/southern Asia to around 8000 km, after which the East Asian populations show an increase towards the Bering land bridge. The two Oceania populations have low to intermediate signals, an probably account for the appearance of a broad double peak in the earlier analyses. Finally, the American populations give a mixed picture, but three of the four have low *t*-values.

In the GLM analysis, motif, fitted as a factor with four levels, is individually the most significant term. Motif type acts to modulate mean t, t increasing in the order GATA, ATT to AC, with the 'others' class, which are predominantly tetranucleotides, appearing most similar to the GATA class. This effect cannot be owing to differences in mean heterozygosity among motifs because heterozygosity is held constant as a separate predictor variable. Instead, it suggests different levels of homoplasy for any given level of heterozygosity, with dinucleotide motifs having the least homoplasy and tetranucleotides the most. Finally, the GLM indicates a significant impact of log modern population size.

4. DISCUSSION
Many studies of human demographic history have concluded that human genetic diversity eroded as we colonized the world from Africa (Harpending & Rogers 2000; Prugnolle *et al.* 2005; Ramachandran *et al.* 2005; Li *et al.* 2008), but very few have attempted to quantify where diversity was lost. The most direct attempt (Hellenthal *et al.* 2008) identifies two putative bottlenecks, one around Africa and one around the Bering land bridge (Fagundes *et al.* 2008), based on step changes in diversity. However, without a correction for population sampling density, this approach tends to identify bottlenecks wherever the coverage of populations is low, owing to an overall pattern of isolation by distance. This method might also be misleading where levels of gene flow vary markedly across the globe. Similarly, analysis of microsatellite data reveals a pattern of steady loss that increases with distance from the Bering Strait, but this
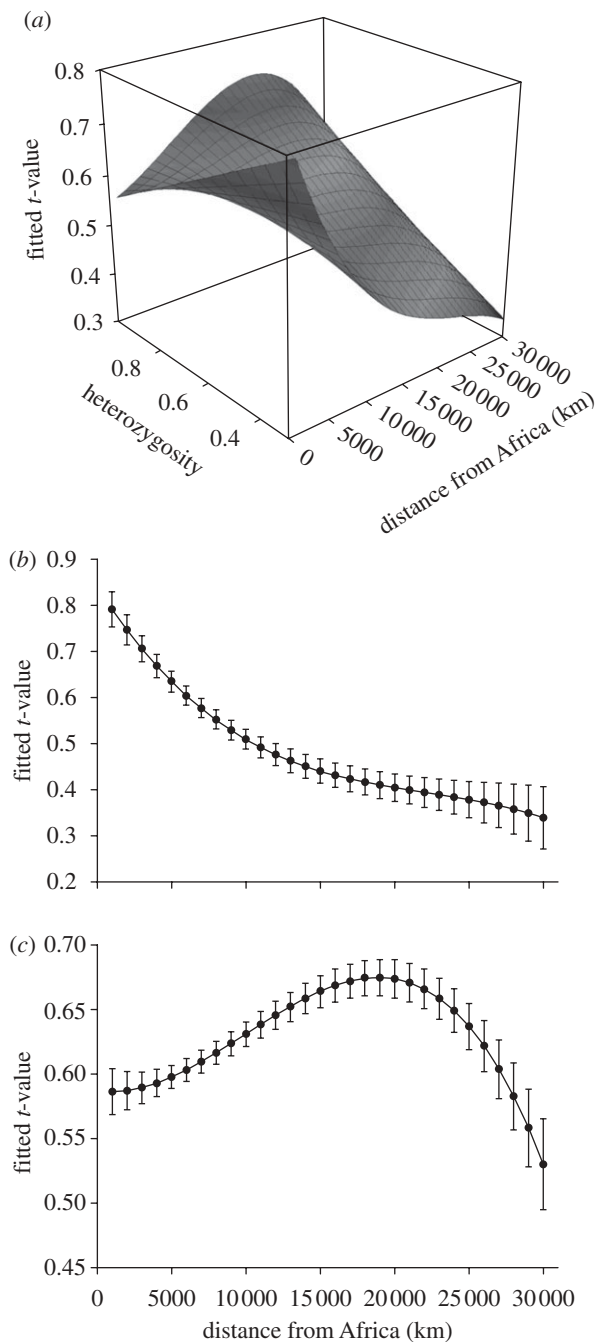
(a)

(b)

(c)

Figure 1. Variation in strength of evidence of a bottleneck (fitted *t*-value) with distance from Africa and heterozygosity in Africa. (*a*) A three-dimensional plot of the fitted values from a general linear model (table 1). (*b*,*c*) Plots of how fitted *t*-values vary with distance from Africa (±s.e.m.) for (*b*) low (heterozygosity = 0.3) and (*c*) high (heterozygosity = 0.9) heterozygosity values. In all cases, the response variable is transformed *t*, and higher values indicate stronger evidence of a bottleneck.
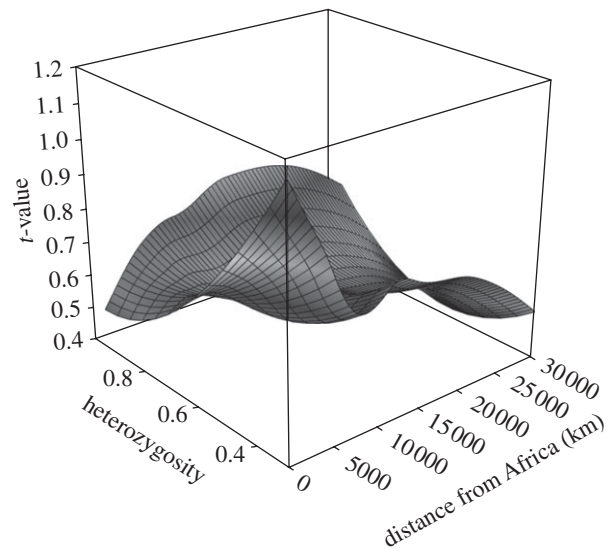

Figure 2. Variation in strength of evidence of a bottleneck (raw data) with distance from Africa and heterozygosity in Africa. This is the same plot as figure 1*a* except that fitted *t*-values have been replaced with the raw data and a surface fitted using local regression. Heterozygosity is taken as that in Africa, as estimated from the Biaka Pygmy sample.

regions with few population samples will contain fewer data, but will not tend to give the impression of bottlenecks simply owing to the larger differences in diversity between adjacent populations. By fitting a flexible GLM to the data, some degree of interpolation is possible, while smoothed splines and individual population values reveal more detailed variation. Our results provide strong support for previous conclusions that human demographic history has featured two different bottlenecks, one close to Africa and one at around 19 000 km away, broadly coincident with the Bering land bridge where humans crossed from Asia into the Americas (Hey 2005; Wang *et al.* 2007). These coincide with the most probable regions where bottlenecks might be expected to have occurred: one as a subset of a larger population moved out of Africa and the other as a further subset of people braved the harsh Arctic terrain to cross a soon-to-be lost land bridge. Unfortunately, the absence of samples from Australia in this dataset means that this part of the expansion could not be examined, even though this population was colonized early (Hudjashov *et al.* 2007), and was isolated by loss of a second land bridge.

Several factors exert a strong influence over the strength of the bottleneck signal. First, mean *t* varies with the mutation model on which the predicted relationship between allele number and heterozygosity is based. The key determinant seems to be the level of homoplasy, with the strict SMM having the most homoplasy and the IAM the least. Greater homoplasy reduces the heterozygosity per allele, making bottlenecks harder to detect. Consequently, the strongest signal is seen using the IAM, even though this model is unrealistic for real microsatellites (Di Rienzo *et al.* 1994; Xu *et al.* 2000). However, while the significance values for the different mutation models differ in their means, they remain highly correlated, indicating that rank significance does not depend
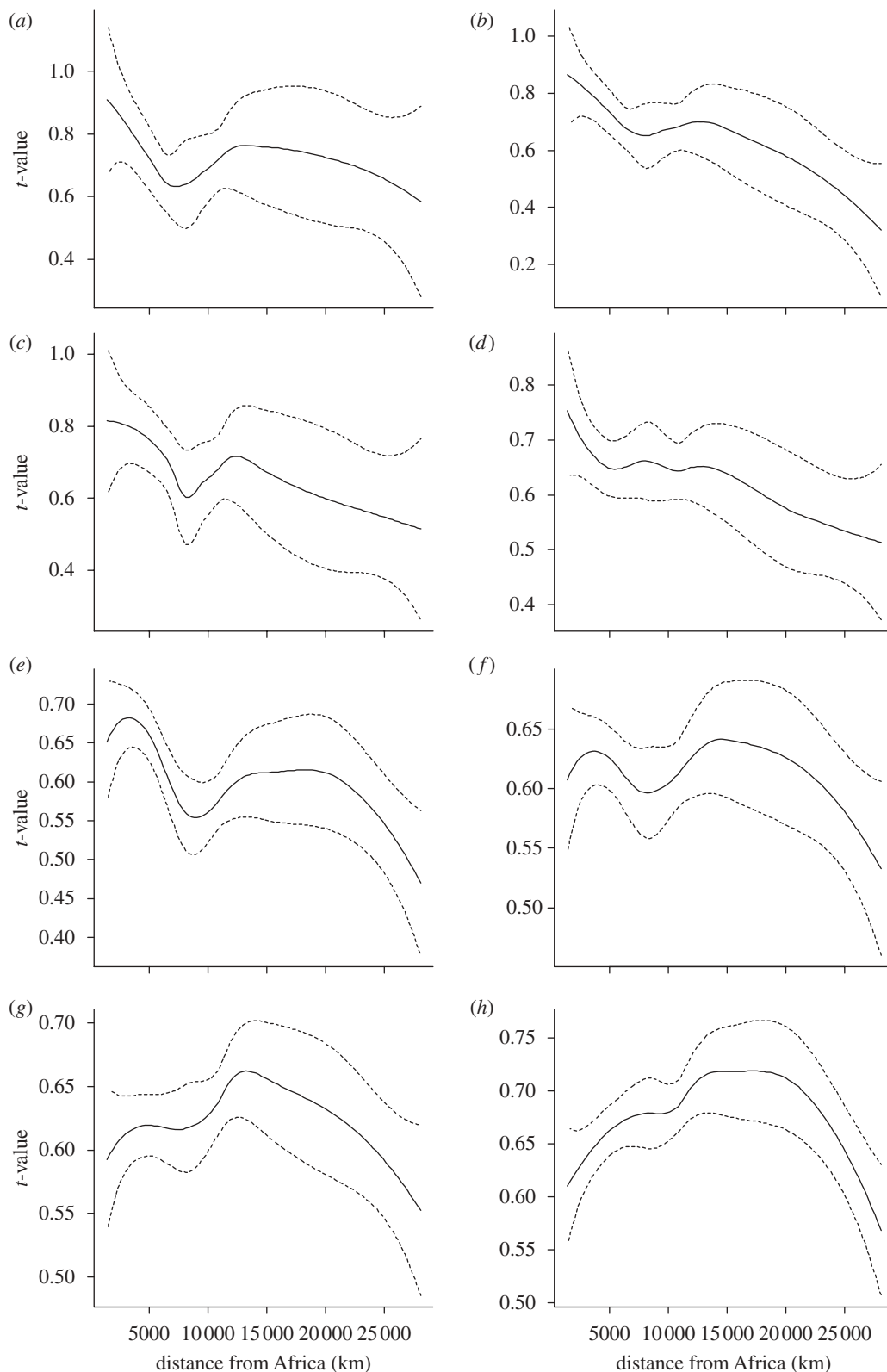
analysis does not test explicitly for a bottleneck (Wang *et al.* 2007). Elsewhere, focus on specific hypotheses such as the peopling of the Americas (Hey 2005) have suggested bottlenecks but at the expense of revealing more global patterns.

Our approach appears novel in that it assesses the evidence for a bottleneck in each population separately, thereby bypassing the need to assume, for example, a uniform pattern of isolation by distance. In this way,

Figure 3. Variation in strength of evidence of a bottleneck (raw data) with distance from (*a*) <0.5 (*n* = 14), (*b*) 0.5−0.55 (*n* = 17), (*c*) 0.55−0.6 (*n* = 13), (*d*) 0.6−0.65 (*n* = 46), (*e*) 0.65−0.7 (*n* = 99), (*f*) 0.7−0.75 (*n* = 162), (*g*) 0.75−0.8 (*n* = 201) and (*h*) >0.8 (*n* = 231). Africa and heterozygosity in Africa. Each panel represents a slice through figure 2 for a different range of heterozygosity values: Bin boundaries were selected to lie 0.05 apart unless this embraced fewer than 10 observations, in which case they were enlarged. In each case local regression is used to fit a smoothed spline and the derive 95 per cent confidence intervals on that spline. The response variable is the transformed *t*-value, higher values indicating stronger evidence of a bottleneck.

on which model is selected. Since we are primarily interested in the relative strength of signal between populations, the choice of mutation model is therefore not critical.

Mean *t* also varies with the heterozygosity of the loci being studied. Some of this appears owing to observation biases, with the lowest variability loci in particular showing a strong dependency on how heterozygosity is
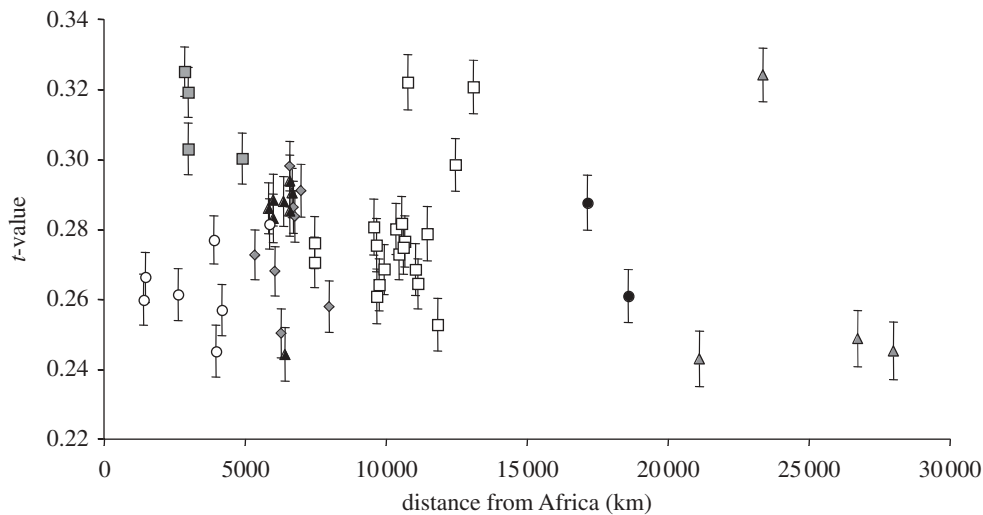
Figure 4. Variation in strength of evidence of a bottleneck with distance from Africa by individual population. Populations are coded: open circles, Africa; grey squares, Middle East; grey diamonds, Europe; black triangles, Central Southern Asia; open squares, East Asia; black circles, Oceania; grey triangles, America. The response variable is the transformed *t*-value, higher values indicating stronger evidence of a bottleneck. Error bars are ± 1 s.e.m.

measured. Despite this, broadly consistent signals can be extracted: low variability loci yield a primary signal of a bottleneck leaving Africa, while high heterozygosity loci yield their strongest peak around the Bering land bridge. We believe the most parsimonious explanation for this relates to the extent to which variability is recovered after the loss that occurred leaving Africa. Low heterozygosity loci have low mutation rates and would have recovered less variability between Africa and the Bering land bridge, potentially reducing the impact of the Bering bottleneck. Conversely, high variability loci may have regenerated the rare alleles that contribute most to the bottleneck signal, at the same time allowing the Bering event to be detected and perhaps reducing the footprint of the out of Africa event. Elucidating more thoroughly the relationship between mutation rate and the timing of events that are detected provides an interesting avenue for future research.

Mean *t* is also influenced by modern population size. This probably reflects the faster rate at which neutral genetic drift operates in small compared with large populations. Populations that became and remained small may be still shedding variability or have reached a new mutation–drift equilibrium. By contrast, populations that became small but re-expanded might have either failed to lose as much variability or, if a strong bottleneck signal was generated, better preserved this signal when population expansion slowed the rate of drift. Thus, while it is easy to see how modern size can influence the bottleneck signal, predicting the direction of the outcome is difficult. Our analysis illustrates how relatively modern demography can impact on our ability to detect historical events.

Finally, mean *t* was influenced by motif type, decreasing in order from AC to ATT and GATA. This is not owing to differences in heterozygosity among motifs but seems to reflect differences in the way the motifs evolve. For a given bottleneck, the strongest predictor of *t* seems to be the degree of homoplasy in the mutation model. Thus, for any given level of heterozygosity, dinucleotide motifs

probably exhibit less homoplasy than equivalent tri- and tetranucleotide motifs. This might be because dinucleotides suffer a higher proportion of larger 'jump' mutations (Di Rienzo *et al.* 1994) or, perhaps, that different motifs vary in their frequency of interruption mutations, which in turn create allelic lineages with widely different rates of slippage (Jin *et al.* 1996; Kruglyak *et al.* 1998). Here, the program BOTTLENECK may provide an interesting new tool for the study of microsatellite evolution.

Despite these complications, a rather consistent pattern emerges, with evidence of a bottleneck being strongest in the Middle East and in the easternmost East Asian/northernmost American populations. These two locations are as one might expect, but there are two additional features that are less obvious. First, the African populations, although at most loci having low *t*-values, do provide quite strong and consistent evidence of a bottleneck at the lowest variability loci. As discussed, this may reflect an observation bias in which loci with very low variability in Africa are unusual for some reason other than demography. An alternative explanation is that these loci still retain the signal of an even more ancient, within-Africa event. This would be consistent with the notion that locus variability is inversely related to the antiquity of the bottleneck signal that is best retained and offers an intriguing hypothesis for future studies. The second feature is the pronounced dip in *t*-value between Europe/central southern Asia and East Asia. This may simply reflect a null signal between two bottlenecks, but might alternatively indicate some other demographic event such as a period of stasis and population expansion. Again, further work is desirable.

In conclusion, we have applied an often-used method for inferring population bottlenecks to worldwide data for human microsatellites. We uncover strong signals of two distinct bottlenecks, one as we left Africa and one as we entered the Americas across the Bering land bridge. By implication, loss of diversity was not as smooth as plots of variability against distance from Africa might suggest.

REFERENCES

Amos, W. & Clarke, A. 2008 Body temperature predicts maximum microsatellite length in mammals. *Biol. Lett.* **4**, 399–401. (doi:10.1098/rsbl.2008.0209)

Betti, L., Balloux, F., Amos, W., Hanihara, T. & Manica, A. 2008 Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proc. R. Soc. B* **276**, 809–814. (doi:10.1098/rspb.2008.1563)

Cavalli-Sforza, L. L., Piazza, A. & Menozzi, P. 1994 *History and geography of human genes.* Princeton, NJ: Princeton University Press.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M. & Slatkin, M. 1994 Mutational processes of simple sequence repeat loci in human populations. *Proc. Natl Acad. Sci. USA* **91**, 3166–3170. (doi:10.1073/pnas.91.8.3166)

Fagundes, N. J. R., Kanitz, R. & Bonatto, S. L. 2008 A re-evaluation of the native American mtDNA genome diversity and its bearing on the models of early colonisation of Beringia. *PLoS ONE* **3**, e3157. (doi:10.1371/journal.pone.0003157)

Harpending, H. & Rogers, A. 2000 Genetic perspectives on human origins and differentiation. *Annu. Rev. Genom. Hum. Genet.* **1**, 361–385. (doi:10.1146/annurev.genom.1.1.361)

Hellenthal, G., Auton, A. & Falush, D. 2008 Inferring human colonization history using a copying model. *PLoS Genet.* **4**, e1000078. (doi:10.1371/journal.pgen.1000078)

Hey, J. 2005 On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol.* **3**, e193. (doi:10.1371/journal.pbio.0030193)

Huang, Q.-Y., Xu, F.-H., Shen, H., Deng, H.-Y., Liu, Y.-J., Liu, Y.-Z., Li, J.-L., Recker, R. R. & Deng, W. 2002 Mutations patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.* **70**, 625–634. (doi:10.1086/338997)

Hudjashov, G. *et al.* 2007 Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc. Natl Acad. Sci. USA* **104**, 8726–8730. (doi:10.1073/pnas.0702928104)

Jin, L., Macaubas, C., Hallmayer, J., Kimura, A. & Mignot, E. 1996 Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *Proc. Natl Acad. Sci. USA* **93**, 15 285–15 288. (doi:10.1073/pnas.93.26.15285)

Kimmel, M., Chakraborty, R., King, J. P., Bamshad, M., Watkins, W. S. & Jorde, L. B. 1998 Signatures of population expansion in microsatellite repeat data. *Genetics* **148**, 1921–1930.

Kimura, M. & Ohta, T. 1978 Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl Acad. Sci. USA* **75**, 2868–2872. (doi:10.1073/pnas.75.6.2868)

Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA* **95**, 10774–10778. (doi:10.1073/pnas.95.18.10774)

Li, J. Z. *et al.* 2008 Worldwide human relationships inferred from genomewide patterns of variation. *Science* **319**, 1100–1104. (doi:10.1126/science.1153717)

Linz, B. *et al.* 2007 An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**, 915–918. (doi:10.1038/nature05562)

Liu, H., Prugnolle, F., Manica, A. & Balloux, F. 2006 A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79**, 230–237. (doi:10.1086/505436)

Luikart, G., Allendorf, F. W., Cornuet, J.-M. & Sherwin, W. B. 1998 Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J. Hered.* **89**, 238–247. (doi:10.1093/jhered/89.3.238)

Manica, A., Amos, W. & Balloux, F. 2007 The effect of ancient population bottlenecks on human phenotypes. *Nature* **448**, 346–348. (doi:10.1038/nature05951)

Prugnolle, F., Manica, A. & Balloux, F. 2005 Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15**, R159–R160. (doi:10.1016/j.cub.2005.02.038)

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. & Cavalli-Sforza, L. L. 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15 942–15 947. (doi:10.1073/pnas.0507611102)

Rogers, A. R. & Harpending, H. 1992 Population growth makes waves in the distribution of pairwise genetic distances. *Mol. Biol. Evol.* **9**, 552–569.

Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. 2003 Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422. (doi:10.1086/380416)

Rosenberg, N. A., Maghajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K. & Feldman, M. W. 2005 Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* **1**, 660–671. (doi:10.1371/journal.pgen.0010070)

Schlötterer, C., Ritter, R., Harr, B. & Brem, G. 1998 High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Mol. Biol. Evol.* **15**, 1269–1274.

Wallace, D. C., Garrison, K. & Knowler, W. C. 1985 Dramatic founder effects in Amerindian mitochondrial DNAs. *Am. J. Phys. Anthropol.* **68**, 149–155. (doi:10.1002/ajpa.1330680202)

Wang, S. *et al.* 2007 Genetic variation and population structure in native Americans. *PLoS Genet.* **3**, e185. (doi:10.1371/journal.pgen.0030185)

Weber, J. L. & Wong, C. 1993 Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128. (doi:10.1093/hmg/2.8.1123)

Xu, X., Peng, M., Fang, Z. & Xu, X. 2000 The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24**, 396–399. (doi:10.1038/74238)