

## ORIGINAL ARTICLE

## How accurate is the current picture of human genetic variation?

IG Romero<sup>1</sup>, A Manica<sup>2</sup>, J Goudet<sup>3</sup>, LL Handley<sup>4</sup> and F Balloux<sup>5</sup><sup>1</sup>Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Cambridge, UK; <sup>2</sup>Department of Zoology, University of Cambridge, Cambridge, UK; <sup>3</sup>Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland; <sup>4</sup>Department of Biological Sciences, The University of Hull, Hull, UK and <sup>5</sup>Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, London, UK

Our understanding of the distribution of worldwide human genomic diversity has greatly increased over recent years thanks to the availability of large data sets derived from short tandem repeats (STRs), insertion deletion polymorphisms (indels) and single nucleotide polymorphisms (SNPs). A concern, however, is that the current picture of worldwide human genomic diversity may be inaccurate because of biases in the selection process of genetic markers (so-called 'ascertainment bias'). To evaluate this problem, we first compared the distribution of genomic diversity between these three types of genetic markers in the populations from the HGDP-CEPH panel for evidence of bias or incongruities. In a second step, using a very relaxed set of

criteria to prevent the intrusion of bias, we developed a new set of unbiased STR markers and compared the results against those from available panels. Contrarily to recent claims, our results show that the STR markers suffer from no discernible bias, and can thus be used as a baseline reference for human genetic diversity and population differentiation. The bias on SNPs is moderate compared to that on the set of indels analysed, which we recommend should be avoided for work describing the distribution of human genetic diversity or making inference on human settlement history.

*Heredity* (2009) **102**, 120–126; doi:10.1038/hdy.2008.89; published online 3 September 2008

**Keywords:** ascertainment bias; microsatellites; SNPs; indels; genetic diversity; population differentiation

## Introduction

The description of worldwide genetic structure of the human population has vastly evolved from the days of ABO blood group typing. The advent of large-scale genomic data has proved crucial to furthering knowledge of the evolution of humankind and subsequent diffusion out of Africa. The recently published HapMap Phase II harbours over three million single nucleotide polymorphisms (SNPs) typed in four distinct human populations (International HapMap Consortium, 2007). The HGDP-CEPH panel (Cann *et al.*, 2002), a resource with a far more extensive geographic coverage, has been analysed extensively, with genotypes at nearly 655 000 genetic markers being freely available (650 000 of these markers are SNPs that have been very recently genotyped (Li *et al.*, 2008)), whereas the remaining loci include short tandem repeats (STRs), and insertion deletion polymorphism (indels). Several follow-up projects, using the STRs and indels previously typed on the HGDP-CEPH samples have added a large number of further populations from targeted geographic regions. These comprise Indians (Rosenberg

*et al.*, 2006), Native Americans (Wang *et al.*, 2007), Polynesians (Friedlaender *et al.*, 2008) and Africans (S Tishkoff, Personal communication).

Many analyses so far have focused on the STR markers from the HGDP-CEPH data set. A general conclusion from these studies was that human genetic variation is greatly influenced by geography, with genetic differentiation between populations increasing with geographic distance and genetic diversity decreasing with distance from Africa (Relethford, 2004; Prugnolle *et al.*, 2005; Ramachandran *et al.*, 2005; Liu *et al.*, 2006). There has also been some debate over the relative importance of the discontinuities in this pattern of isolation by distance (Kittles and Weiss, 2003; Serre and Pääbo, 2004; Manica *et al.*, 2005; Rosenberg *et al.*, 2005; Witherspoon *et al.*, 2006; Handley *et al.*, 2007; Linz *et al.*, 2007). Recent developments in marker discovery will undoubtedly allow this picture to be refined and lead to a much-improved understanding of human population structure and past settlement history. This wealth of data should also prove crucial to help characterizing regions of the genome under natural selection (for example, Sabeti *et al.*, 2006), by providing baseline linkage patterns, population diversity and differentiation. Yet, there remains a nagging concern about the accuracy of the emerging picture of the apportionment of human genomic diversity. The problem is that genetic markers are generally not selected as a random subset of all the available ones, but represent a subset that satisfies specific criteria, in particular prior knowledge of them having high genetic diversity.

Correspondence: Dr F Balloux, Department of Infectious Disease Epidemiology, MRC Centre for Outbreak Analysis and Modelling, Imperial College Faculty of Medicine, St Mary's Campus, Norfolk Place, London W2 1PG, UK or Dr A Manica, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK.  
E-mails: fballoux@imperial.ac.uk or am315@cam.ac.uk

Received 7 April 2008; revised 24 June 2008; accepted 11 July 2008; published online 3 September 2008

Locus discovery procedures have been criticized before as sources of insidious bias, a problem described as ascertainment bias particularly recognized for SNPs (Rogers and Jorde, 1996; Kuhner *et al.*, 2000; Wakeley *et al.*, 2001; Akey *et al.*, 2003; Bustamante *et al.*, 2005). DNA sample panels used for marker discovery and initial screening are generally small compared to the final sample and do not contain members of all populations being studied. Such a procedure will preferentially capture polymorphisms represented in the discovery panel. European subpanels have been frequently used in locus discovery procedures, and it has already been argued that they could be inflating existing estimates of human diversity towards Europe, primarily at the expense of African variation (Carlson *et al.*, 2003; Tishkoff and Verreli, 2003). STRs have been traditionally deemed to be immune from ascertainment bias thanks to their high mutation rate (Bowcock *et al.*, 1994; Rogers and Jorde, 1996; Harpending and Rogers, 2000). However, two recent studies have concluded that the set of STRs deployed on the HGDP-CEPH panel were suffering from serious ascertainment bias (Ray *et al.*, 2005; Foll and Gaggiotti, 2006).

Given that these panels already have, and will continue to represent the material for a substantial fraction of inference on human population genetics, we felt compelled to examine whether ascertainment bias may be a confounding factor. To address this, we examined three different subsets of markers previously genotyped in the HGDP-CEPH panel, including the 783 STR and 210 indel data sets generated by Rosenberg *et al.* (2005) and 2834 SNPs typed by Conrad *et al.* (2006). We also developed a new set of STRs, which were designed using very relaxed criteria and tested without screening for high polymorphism, and genotyped them in the HGDP-CEPH panel. This unbiased marker set provides a null control, against which, published data sets can be assessed for the possibility of ascertainment bias.

## Materials and methods

### Data sources

Individuals were split into the 54 HGDP-CEPH populations for all analyses, treating Bantu SW and Bantu SE individuals as separate populations. Analysis of the STRs and indel data sets was performed on the H971 subset of the HGDP-CEPH panel (Rosenberg, 2006), which includes suspected second-degree relative pairs, but no first-degree relative pairs. The SNP data set was originally genotyped in a different subset of the HGDP-CEPH panel, H927, which contains no first- or second-degree relatives, and also excludes 22 individuals with large amounts of missing data.

### Design of new loci

To define the new panel of 'unbiased STRs', 70 STR loci (Supplementary Table 1) were designed according to the following criteria: (1) loci must be unlinked, (2) loci must be tetranucleotides, (3) STR sequence must contain at least four perfect consecutive repeats, (4) STR sequence must contain between 5 and 50 repeats in total, (5) there must be no extra repeats between the main STR sequence and the primer binding site and (6) it must be possible to design a specific primer pair for the locus yielding a

product no bigger than 500 bp. We limited the choice to tetranucleotides for their ease of scoring and because they represent the dominant repeat type (71%) in the STRs already genotyped in the HGDP-CEPH panel.

Upwards to 500 000 bp from a randomly selected contig were retrieved from the human genome reference sequences (loci Cam001–Cam040, Build 34; loci Cam041–Cam070, Build 35) and analysed for the presence of STRs using the program Tandem Repeat Finder (Benson, 1999). To account for the possibility of linkage, only one contig was chosen from each chromosomal arm; for acrocentric chromosomes contigs were selected from the centromeric and telomeric ends of the q arm (Supplementary Table 1). The fifth identified STR locus in each contig was used, provided it fulfilled the criteria outlined above. Where it did not, the sixth locus was examined, and so on until an acceptable locus was found in each retrieved contig.

Primers for the identified loci were designed using Primer3 (Rozen and Skaletsky, 2000) and optimized on eight individuals of diverse geographic background in a standard 10  $\mu$ l PCR reaction using a Touchdown protocol (see Supplementary Table 2 for primer and reaction conditions). A total of 36 primer pairs yielded products of the expected size and 5'-fluorescent-labelled primers were ordered for these loci and tested in a single plate of the HGDP-CEPH panel. Six loci subsequently failed fluorescent amplification. The remaining 30 loci were amplified in the H971 subset of the HGDP-CEPH panel, genotyped offsite on an ABI 3730 and scored manually using GeneMapper 3.7 (Applied Biosystems, Foster City, CA, USA). Of the 30 loci, another six were discarded because they proved impossible to score. In total, 24 randomly designed loci were genotyped in the entire HGDP-CEPH panel. Eight essentially invariant hence uninformative loci ( $H_S < 0.01$ ) were excluded from all subsequent analyses, resulting in a final set of 16 'unbiased' polymorphic loci (full genotypes in Supplementary Table 3).

### Statistical analyses

Genetic diversity ( $H_S$ ) and matrices of pairwise population differentiation ( $F_{ST}$ ) were calculated using FSTAT 2.9.3 (Goudet, 2002). The test for origin, slope and variance explained ( $R^2$ ) between the new loci and the previous 789 STRs were obtained by a resampling approach. We generated 1000 data sets by resampling 16 loci out of the 789 available. We calculated population genetic diversities ( $H_S$ ) and all pairwise genetic distances ( $F_{ST}$ ) on every resampled data set, and these were then correlated to distance from Africa and pairwise geographic distances along landmasses respectively.  $P$ -values were obtained directly from the ranking of the intercept, slope or variance explained at the new 16 loci in the distribution of the 1000 correlations obtained on resampled data sets.

### Geographic analyses

Geographic distances computed as shortest distances along landmasses and avoiding mountain regions with average altitude  $>2000$  m were computed within a spherical referential using graph theory (Manica *et al.*, 2005; Prugnolle *et al.*, 2005). The hypothetical origin for anatomically modern humans was chosen as  $-12^\circ$

latitude and 25° longitude (Manica *et al.*, 2007). We also considered an East African origin (Addis Ababa; latitude 9°, longitude 38.4°) for the correlation shown in Figures 2a–c; the alternative origin did not affect the results qualitatively. The interpolations in Figure 2 were obtained with an inverse distance-weighted (IDW) interpolation method on landmasses using the ArcGIS Spatial Analyst extension.

## Results

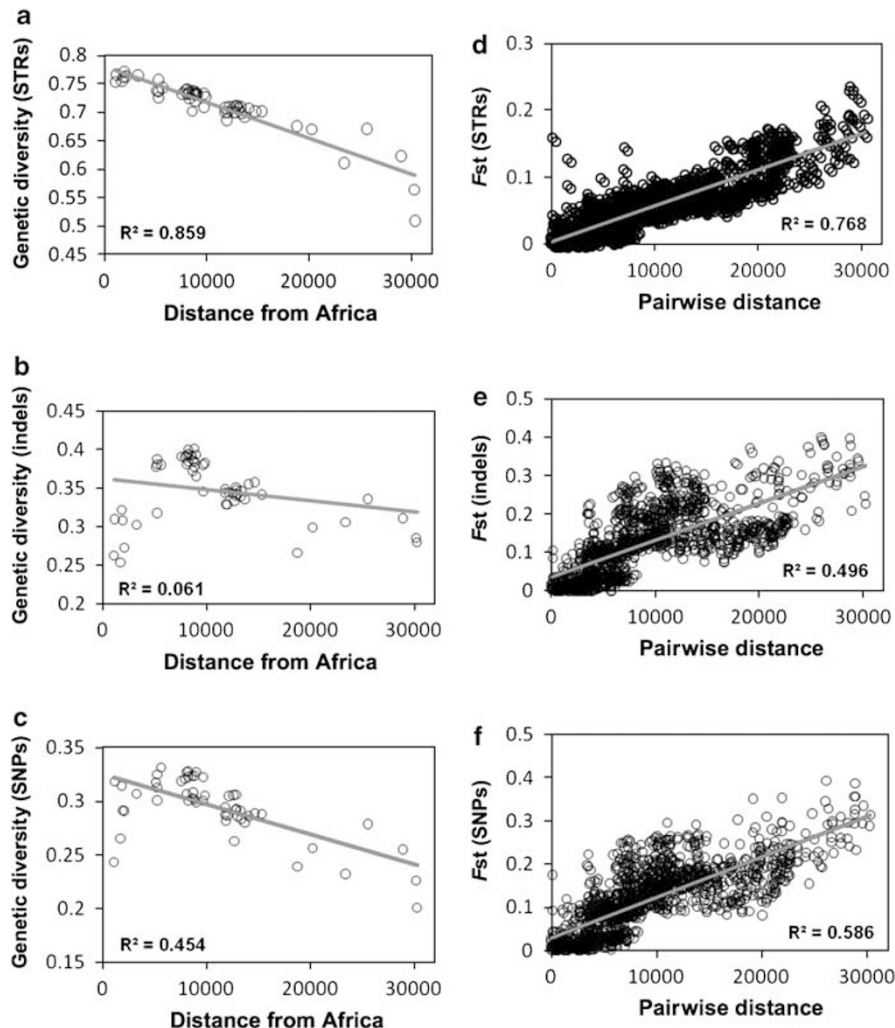
### Patterns in genetic diversity of available markers

To contrast the different classes of markers we first plotted genetic diversity as a function of geographic distance along landmasses from a hypothetical African origin. Figure 1 (panels a–c) shows plots of mean population genetic diversity ( $H_S$ ) for the 54 HGDP-CEPH populations against geographic distance along landmasses from an African centroid (Manica *et al.*, 2007) for the three sets of markers. The strongest correlation between distance and genetic diversity is shown by the 783 STRs ( $R^2 = 0.859$ ), but SNPs also show a moderate

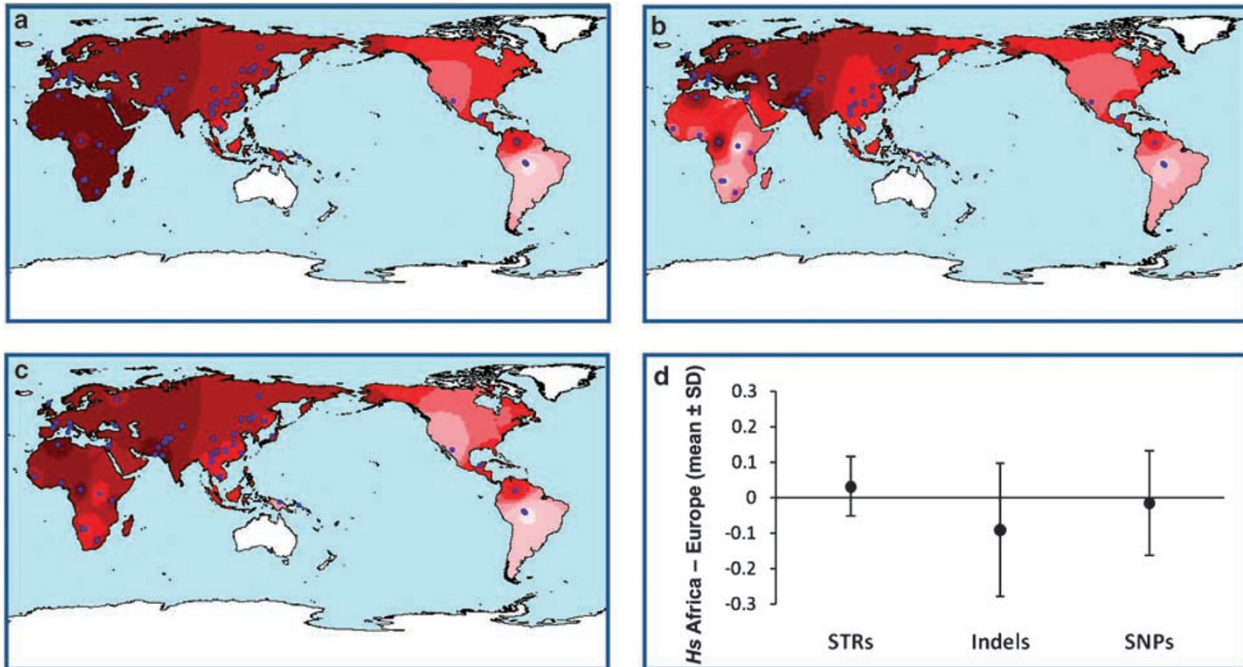
correlation ( $R^2 = 0.454$ ). The STR and SNP data sets therefore corroborate previous observations of a loss of genetic diversity with increasing distance from Africa. By contrast, indels show a very poor correlation ( $R^2 = 0.061$ ). European and Pakistani samples in the indel data set show markedly higher diversity than expected given their distance from Africa, and compared to African samples, which are themselves scattered far below the cline.

In a second step we consider how well different classes of markers fit within an isolation by distance model. Plots of pairwise population genetic differentiation ( $F_{ST}$ ) against geographic distance along landmasses are shown in Figures 1d–f. STRs display the strongest pattern of isolation by distance ( $R^2 = 0.768$ ) whereas SNPs correlate moderately well ( $R^2 = 0.586$ ). In this case, the indels correlate much better than before ( $R^2 = 0.496$ ) but nonetheless visibly stray from linearity, with in particular high  $F_{ST}$  values for pairs involving one African and one Eurasian population.

To get a closer look at the distribution of genetic diversity we interpolated genetic diversities for the 54 population for the three classes of markers



**Figure 1** The panels on the left-hand side show genetic diversities against distance from Africa along landmasses for short tandem repeats (STRs) (a), indels (b) and single nucleotide polymorphisms (SNPs) (c). Panels on the right-hand side show pairwise geographic distance along landmasses against pairwise genetic distance ( $F_{ST}$ ) for STRs (d), indels (e) and SNPs (f).



**Figure 2** Interpolation of estimates of genetic diversity ( $H_s$ ) for short tandem repeats (STRs) (a), indels (b) and single nucleotide polymorphisms (SNPs) (c). The intensity of the red colour represents the genetic diversity obtained with an inverse distance-weighted (IDW) interpolation on landmasses. Blue dots represent the 54 populations from the H971 subset of the HGDP-CEPH data set. (d) The difference in genetic diversity between African and European populations for the three classes of markers. Error bars report standard deviation.

(Figures 2a–c). For STRs, the highest diversity is found in Africa, a result in line with our understanding of human population genetics (Bowcock *et al.*, 1994; Harpending and Rogers, 2000). This set of indels on the other hand display an intriguing pattern with a hotspot of diversity in Europe and also substantial polymorphism in East Asia. There is a decrease around India and Pakistan and even lower values for the African continent. SNPs show a pattern somewhere in between the two previous classes of markers, with a maximum in Europe, but slightly lower diversities in Africa.

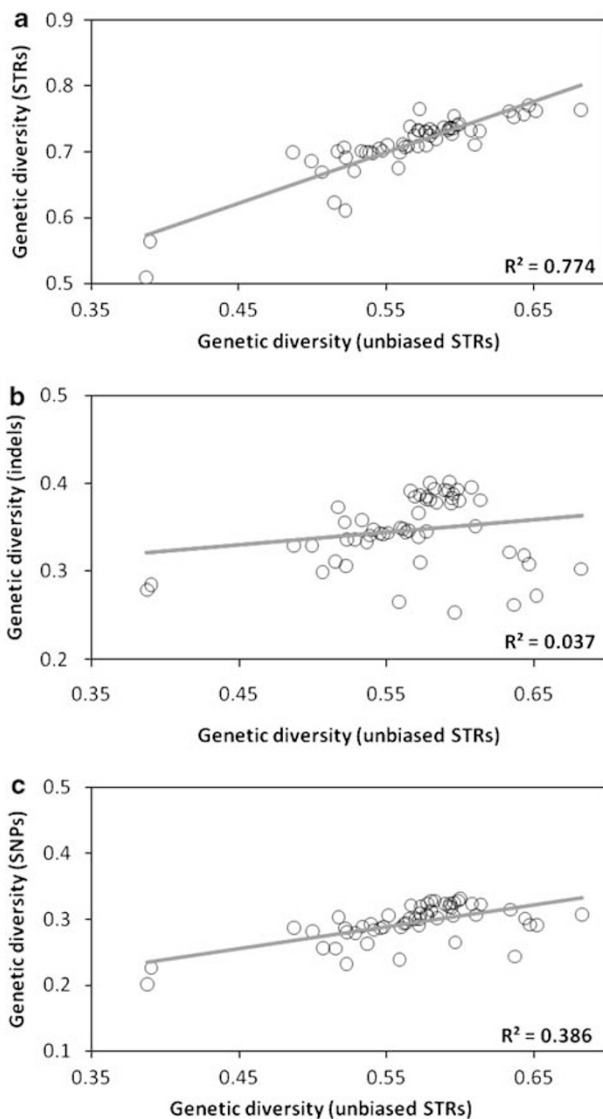
To get a more quantitative feel for the geographic distribution of genetic diversity we chose to contrast in Figure 2d the difference in average genetic diversity of African populations and European populations (average of African populations–average of European populations). Genetic diversity over loci is significantly higher in Africans than Europeans for STRs (paired  $t$  (782) = 10.37,  $P < 10^{-23}$ ), whereas the reverse is true for indels (paired  $t$  (209) = -7.08,  $P < 10^{-10}$ ) and SNPs (paired  $t$  (2833) = -4.21,  $P = 2.6 \times 10^{-5}$ ). To highlight the spread over individual loci irrespective of sample size, we report standard deviation in the error bars, which illustrate the more homogeneous behaviour of the STR loci compared to the SNPs and indels.

#### Comparison with the new unbiased markers

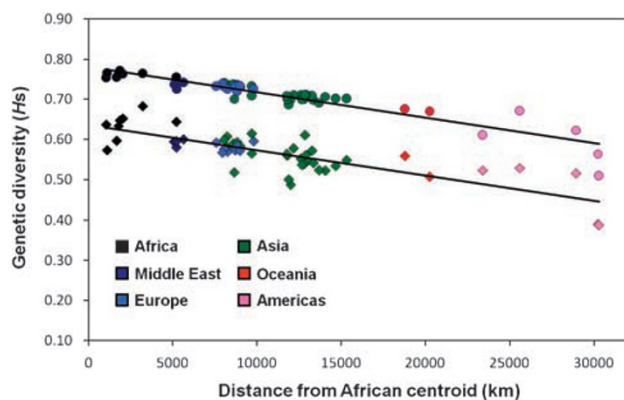
Although we can conclude at this stage that indels and SNPs seem affected by ascertainment bias, we need an external ‘unbiased control’ to reach any statement about the STRs. Thus, we designed a new set of ‘unbiased STRs’ and genotyped them in the HGDP-CEPH panel. In a first step, we explored the correlation over the 54

populations between the genetic diversity of the new panel of STRs against previous STRs, indels and SNPs (Figure 3). Although the correlation between the two sets of STRs is respectable ( $R^2 = 0.774$ ), the correlation with the SNPs is considerably lower ( $R^2 = 0.386$ ) and hardly different from zero for the indels ( $R^2 = 0.034$ ). The new loci are less variable in all populations than the loci included in the 783 STR panel (intercept  $P \leq 0.001$ ). This is to be expected as the unbiased loci were not selected for their high polymorphism. The new STRs are nevertheless more variable than either the indels or the SNPs.

There is a strong correlation between mean genetic diversity ( $H_s$ ) per population and distance from sub-Saharan Africa (Figure 4, lower line, diamonds,  $R^2 = 0.611$ ). The slope of the decrease in intrapopulation variation over geographic distance is essentially identical to that of the 783 STRs examined above ( $-6.29 \times 10^{-6}$  and  $-6.32 \times 10^{-6}$  for the new and previous STRs respectively,  $P = 0.922$ ). The mean correlation over the 1000 resampling runs ( $R^2 = 0.694$ ) is not significantly different from the one obtained on the 16 new STRs ( $P = 0.521$ ). Although the new set of ‘unbiased STRs’ are far less diverse, this reduction in polymorphism is highly consistent across populations, and we could detect no evidence for ascertainment bias in the 783 STR data set. When we compared the isolation by distance patterns between the two sets of markers, we could find no difference in intercept or slope ( $P = 0.398$  and  $0.574$  respectively). There was a non-significant trend for the new markers to explain less variance than a random set of 16 previous STRs ( $P = 0.058$ ). This is most likely due to the lower genetic diversity of the new markers and this trend disappears if we subsample among the 100 least polymorphic loci in the HGDP-CEPH STRs ( $P = 0.280$ ).



**Figure 3** Correlation over 54 populations of the genetic diversity between the new set of ‘unbiased short tandem repeats (STRs)’ against previous STRs (a), indels (b) and single nucleotide polymorphisms (SNPs) (c).



**Figure 4** Genetic diversity ( $H_s$ ) and distance from sub-Saharan Africa for the 783 previous short tandem repeats (STRs; circles, upper line) and the new panel of 16 ‘unbiased loci’ (diamonds, lower line).

## Discussion

Over recent years, our understanding of the distribution of human genomic diversity has progressed dramatically with the availability of large data sets typed at hundreds of genetic markers. However, this picture of the distribution of human genetic diversity and any inference based upon will be limited by the quality of the underlying data. This prompted us to evaluate the biases affecting three classes of markers previously typed on the HGDP-CEPH panel. We confirm the bias affecting these SNPs (Conrad *et al.*, 2006), and the underestimation of diversity in Africans, but could find no evidence for a significant ascertainment bias affecting STRs. By contrast, we did not anticipate the extremely poor behaviour of indels.

### Ascertainment bias in STRs

STRs have been deemed to be essentially immune to ascertainment bias due to their high mutation rate (Bowcock *et al.*, 1994; Rogers and Jorde, 1996; Harpending and Rogers, 2000). This view has been challenged by two recent papers analysing the HGDP-CEPH data set (Ray *et al.*, 2005; Foll and Gaggiotti, 2006). The two papers reached this conclusion through different means but both invoked strong ascertainment bias. Ray *et al.*'s explicit simulation model pointed to an unexpected North African origin of modern humans unless they simulated loci affected by strong ascertainment bias. Ascertainment bias on the available STRs is an unlikely explanation for the shift in the inferred origin, particularly because we could detect no difference in slope or intercept in the isolation by distance pattern between these markers and the new unbiased ones. Foll and Gaggiotti (2006) estimated differentiation ( $F_{ST}$ ) in turn for each population from the HGDP-CEPH panel against all others and plotted these values against distance from Africa. They recovered surprisingly high differentiation for African populations, which they accounted for in part as due to ascertainment bias. Our results indicate ascertainment bias is unlikely to be the main culprit, if at all.

It is difficult to point to the cause of the shift in origin in Ray *et al.* (2005) paper, as their sophisticated simulation framework is highly complex and parameter rich. The methodology used by Foll and Gaggiotti (2006) is more transparent. Their approach was to compute sequentially an  $F_{ST}$  value for each population against all others, which is essentially equivalent to averaging the  $F_{ST}$  value of all pairs involving the focal population. Such an approach is expected to be sensitive to geographic heterogeneity in the distribution of samples. Populations that are part of a big cluster will have lower  $F_{ST}$  values than more geographically isolated populations, because the average  $F_{ST}$  for a population in a cluster will include all near-zero values of the comparisons with populations from the same cluster. The distribution of the HGDP-CEPH populations is highly overdispersed (see Figure 1), with in particular a large number of samples clustered in Europe, around Pakistan and in East Asia, whereas African samples are more evenly distributed.

Our new panel of ‘unbiased markers’ is arguably relatively small and it may miss subtle effects. However, as there is no trend towards significance for either a difference with the previous STRs in the clinal pattern of



decrease in genetic diversity (Figure 4;  $P=0.521$ ) or the relative genetic diversity of Africa versus Europe ( $P=0.173$ ), ascertainment bias in the previous STRs—if there was any—seems far too small to explain these observations. The 783 STR set thus provides the best current resource for baseline genetic diversity and population differentiation at a worldwide scale.

### Strong bias on indels

The extreme bias affecting this set of indels—which are a subset of over 2000 indels described by Weber *et al.* (2002)—came as a surprise. It is doubtful this bias can be fully ascribed to standard ascertainment bias (that is bias for markers which are more polymorphic than average in the sample used for marker discovery). We suspect these loci have been further selected for their high difference in allele frequencies between Africans and Europeans, a desirable property in admixture mapping in African Americans. The extreme underestimation of African genetic diversity and the high differentiation between African and Eurasian populations do not make them suitable for inferences on past human demography at a large geographic scale. This set of markers has not been widely used so far and then only in conjunction with STRs (Ramachandran *et al.*, 2005; Rosenberg *et al.*, 2006; Friedlaender *et al.*, 2008). Because of their far lower information content compared to STRs (information content is proportional to polymorphism in most population genetics analyses), they probably only contributed marginally in the conclusions reached. Yet, given the extent of the bias affecting them we clearly recommend against including them in future work.

### A benchmark for correcting bias in SNP data sets

The bias on this set of SNPs is less severe and has previously been discussed by Conrad *et al.* (2006). Several strategies to alleviate this problem have already been proposed either through the use of fixed window sizes (Conrad *et al.*, 2006) or haplotypes of phased data (Li *et al.*, 2008). The evidence we provide here that the STRs can be considered as essentially bias free renders such approaches even more attractive. The set of SNPs we analysed here has been ascertained from different sources (Perlegen multiethnic panel of 20 individuals, HapMap Phase I and dbSNP double hits). The larger data sets produced on commercial SNP chips will be characterized by different biases. However, the fact that the available STRs can be considered essentially free of bias provides with a benchmark against which various debiasing strategies can be assessed, for instance by evaluating the window size that best restores linearity in the correlation between population genetic diversities at SNPs and STRs. Such an approach should be applicable to essentially any future high-throughput data set generated on the HGDP-CEPH panel, thus making such large data sets more amenable to inference on human settlement history.

### Acknowledgements

We acknowledge support from the BBSRC. The work was performed while JG was on a sabbatical fellowship in Cambridge made possible thanks to the support of an Underwood fellowship administered by the BBSRC. We thank Laurent Excoffier and Matthieu Foll for discus-

sions and comments on the paper. We are also grateful for constructive comments by two anonymous reviewers.

### References

- Akey JM, Zhang K, Xiong M, Jin L (2003). The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol Biol Evol* **20**: 232–242.
- Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.
- Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455–457.
- Bustamante CD, Fedel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S *et al.* (2005). Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153–1157.
- Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L *et al.* (2002). A human genome diversity cell line panel. *Science* **296**: 261–262.
- Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA *et al.* (2003). Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* **33**: 518–521.
- Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA *et al.* (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**: 1251–1260.
- Foll M, Gaggiotti OE (2006). Identifying the environmental factors that determine the genetic structure of populations. *Genetics* **174**: 875–891.
- Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK *et al.* (2008). The genetic structure of Pacific Islanders. *PLoS Genetics* **4**: e19.
- Goudet J (2002). FSTAT, a program to estimate and test gene diversities and fixation indexes (version 2.9.3) (1995). *J Hered* **86**: 485–486.
- Handley LJJ, Manica A, Goudet J, Balloux F (2007). Going the distance: human population genetics in a clinal world. *Trends Genet* **23**: 432–439.
- Harpending H, Rogers A (2000). Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet* **1**: 361–385.
- International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.
- Kittles RA, Weiss KM (2003). Race, ancestry, and genes: implications for defining disease risk. *Annu Rev Genomics Hum Genet* **4**: 33–67.
- Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000). Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**: 439–447.
- Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S *et al.* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–1104.
- Linz L, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P *et al.* (2007). An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **445**: 915–918.
- 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.
- Manica A, Amos W, Balloux F, Hanihara T (2007). The effect of ancient bottlenecks on human phenotypic variation. *Nature* **448**: 346–348.
- Manica A, Prugnolle F, Balloux F (2005). Geography is a better determinant of human genetic differentiation than ethnicity. *Hum Genet* **118**: 366–371.

- Prugnolle F, Manica A, Balloux F (2005). Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15**: R159–R160.
- Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (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**: 15942–15947.
- Ray N, Currat M, Berthier P, Excoffier L (2005). Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res* **15**: 1161–1167.
- Relethford JH (2004). Global patterns of isolation by distance based on genetic and morphological data. *Human Biol* **76**: 499–513.
- Rogers AR, Jorde LB (1996). Ascertainment bias in estimates of average heterozygosity. *Am J Hum Genet* **58**: 1033–1041.
- Rosenberg NA (2006). Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet* **70**: 841–847.
- Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MGB, Nino-Rosales L, Nisus V *et al.* (2006). Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet* **2**: e215.
- Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* **1**: e70.
- Rozen S, Skaletsky H (2000). Primer3 on the www for general users and for biologist programmers. *Methods Mol Biol* **132**: 365–386.
- Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O *et al.* (2006). Positive natural selection in the human lineage. *Science* **312**: 1614–1620.
- Serre D, Pääbo S (2004). Evidence for gradients of human genetic diversity within and among continents. *Genome Res* **14**: 1679–1685.
- Tishkoff S, Verrelli B (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* **4**: 293–340.
- Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001). The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *Am J Hum Genet* **69**: 1332–1347.
- Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G *et al.* (2007). Genetic variation and population structure in Native Americans. *PLoS Genet* **3**: e185.
- Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G (2002). Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet* **71**: 854–862.
- Witherspoon DJ, Marchani EE, Watkins WS, Ostler CT, Wooding SP, Anders BA *et al.* (2006). Human population genetic structure and diversity inferred from polymorphic L1 (*LINE-1*) and *Alu* insertions. *Hum Hered* **62**: 30–46.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)