

50. Helfman, G., Collette, B. & Facey, B. *The Diversity of Fishes* (Blackwell Science, Malden, Massachusetts, 1997).
51. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
52. Katagiri, T. *et al.* Construction and characterization of BAC libraries for three fish species; rainbow trout, carp and tilapia. *Anim. Genet.* **32**, 200–204 (2001).
53. Nelson, J. S. *Fishes of the world* (John Wiley and Sons, New York, 1994).

Acknowledgements

We thank M. Berenbrink for helpful discussions and comments and anonymous referees for comments. A.R.C. was supported by long-term funding from the UK Natural Environmental Research Council who also have supported the Liverpool Microarray Facility. D.L.C. was supported by grants from the US National Science Foundation Biocomplexity Programme and the US National Heart Lung and Blood Institute.

Competing interests statement
The authors declare no competing financial interests.

Online links

FURTHER INFORMATION

dbEST — Database of Expressed Sequence Tags: http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html
Fishbase: www.fishbase.org
Gene Ontology: <http://www.geneontology.org>
KEGG — Kyoto Encyclopedia of Genes and Genomes: www.genome.jp/kegg
Medaka Genome Project: <http://dolphin.lab.nig.ac.jp/medaka>
Tetraodon Genome Browser: www.genoscope.cns.fr/externe/tetraodon
The Danio rerio Sequencing Project: www.sanger.ac.uk/Projects/D_rerio
The Fugu Genomics Project: <http://fugu.hgmp.mrc.ac.uk>
Zebrafish Gene Collection: <http://zgc.nci.nih.gov>
Access to this interactive links box is free online.

OPINION

The Human Genome Diversity Project: past, present and future

L. Luca Cavalli-Sforza

Abstract | The Human Genome Project, in accomplishing its goal of sequencing one human genome, heralded a new era of research, a component of which is the systematic study of human genetic variation. Despite delays, the Human Genome Diversity Project has started to make progress in understanding the patterns of this variation and its causes, and also promises to provide important information for biomedical studies.

The **Human Genome Diversity Project** (HGDP) provides a resource that is aimed at promoting worldwide research on human genetic diversity, with the ultimate goal of understanding how and when patterns of diversity were formed. It also has the added benefit of providing information that is likely to prove useful to several areas of biomedical research. Here, I provide an update on the HGDP, focusing on important progress since earlier reviews¹, the present status of the project² and how this resource could be developed most effectively in the future. I also discuss possible relations with the **International HapMap Project**^{3,4}, another large-scale study of human genome diversity. Despite having generally different aims, the two projects provide complementary resources, indicating that interactions between the two could prove mutually beneficial. Finally, I summarize some desirable future developments for the HGDP

and its potential to improve the understanding of the genetic structure of the human species and facilitate medical applications.

The HGDP — rationale and history

Founding of the HGDP. As early as the beginning of the twentieth century, the potential for genetic data to provide information on the history and geography of human populations was known from the study of proteins (BOX 1). However, until recently, the collection of such data remained largely a piecemeal endeavour. Indeed, it was not until the **Human Genome Project** (HGP) was in full swing that the idea of a large-scale systematic study of human genomic variation was raised⁵. Specifically, it was realized that renewable samples from well-chosen populations, for which any part of the genome could be examined, could greatly facilitate studies of the genetic geography and history of our species.

The foundation of the HGDP was prompted by discussions among geneticists interested in human evolution and population genetics⁶. The President of the **Human Genome Organisation** (HUGO) at that time, Sir Walter Bodmer, asked me to chair a committee to study the feasibility of a human genomic variation project. As the idea developed, this was named the Human Genome Diversity Project. The US National Institutes of Health (NIH) Institute for General Medical Science, the US National Science Foundation, and initially also

the US Department of Energy (which had previously financed mutation rate studies by my group), supported four symposia, between 1991 and 1994, that addressed the genetic and statistical issues, anthropological issues, general organization, and molecular and ethical issues related to the HGDP.

Overcoming initial difficulties. Political and ethical difficulties arose¹ in 1994, similar to those that marked the beginning of the HGP, but in the case of the HGDP they focused especially on the fear that indigenous people might be exploited by the use of their DNA for commercial purposes ('bio-piracy'). However, since its initiation, the HGDP has avoided commercial interests, and when the project was finally ready to be launched, it was made clear that DNA samples would be provided only to non-profit-making laboratories. The HGDP has always opposed the patenting of DNA, to allow the study of genetic variation for fundamental research purposes. Concern that HGDP data would feed 'scientific racism' was also expressed by naive observers, despite the fact that half a century of research into human variation has supported the opposite point of view — that there is no scientific basis for racism. Consequently, agencies that had financed the HGDP organizational symposia asked the US **National Research Council** (NRC) of the National Academy of Sciences (NAS) to convene a committee to study the feasibility and ethics of the project, similar to the evaluation of the HGP at a comparable time in its history.

From 1994 to 1997, while the NRC committee was organized, met and wrote its report, the HGDP took no major action. Instead, it prepared for the final stage of organization, in the hope and expectation that the NRC committee would give a positive response. In particular, at the first Cold Spring Harbor (CSH) Symposium on Human Evolution, held in October 1997, there was a meeting of several research workers who had collected cell lines from indigenous populations. At my request, they unanimously agreed to contribute cell lines to a central collection that would form the core of the HGDP.

Since its initiation, the organizers of the HGDP were convinced that the crucial first effort was to establish a collection of **LYMPHOBLASTOID CELL LINES** (LCLs) from many populations — rather than simply collecting DNA samples — for reasons of accuracy and renewability (BOX 2). The fact that LCLs had already been made from worldwide populations by researchers of human evolution also supported the validity of the approach, and

the donation of these lines to the HGDP made immediate funding unnecessary.

Uncertainties concerning such issues as strategies for collecting samples in a way that would facilitate anthropological or medical research, and the choice of the populations¹, were obviated by the sources of the cell lines, which were donated by researchers working on human evolution. However, as discussed later, it became clear that orientation of the collection towards anthropological interests also offered excellent chances of aiding medical research.

The recommendation of the NAS-NRC committee, made public⁷ at the end of 1997, was that the HGDP could proceed, with particular attention being paid to informed consent and related ethical issues. The NIH Institute of General Medical Sciences, a chief supporter throughout, has constantly followed and revised the ethical rules of the endeavour (BOX 3).

Two important problems had to be solved at the time when the NAS-NRC authorization finally came. The first was the question of where to house the HGDP collection. The Center for the Study of Human Polymorphism (CEPH) at the **Fondation Jean Dausset** in Paris agreed to house and distribute the collection. In 1984, the CEPH had initiated the international collaboration to genetically map the human genome⁸, which was built around a resource of LCLs from 40 large kindreds, and so had all the facilities needed for storing cell lines and distributing large numbers of DNA samples.

The second question concerned whether the collection was adequate for the intended research purposes. This could be decided only after all the LCLs had arrived in Paris. All five continents are represented in the collection, and all samples are from populations of anthropological interest — that is, those that were in place before the great diasporas started in the fifteenth and sixteenth centuries, when navigation of the oceans became possible. This choice was important, because these diasporas caused significant population admixtures, especially in the Americas but also in other continents. Only genetic knowledge of the original populations that contributed to these admixtures can disentangle the various genetic complexities that resulted, and the HGDP fulfils these criteria.

The HGDP collection was to include more than 1,000 cell lines; inevitably, there would be large gaps given the collection strategy. However, it was important to begin the project, as determining the success of the initial collection was thought to be essential for understanding if and how it would be worth expanding.

Box 1 | Some general principles of human genetic diversity studies

Gene-based studies of genetic variation

Data on human genetic variation, collected since 1919 on proteins^{31–34}, and more recently on DNA, have been widely used for the reconstruction of human history^{9,15,25}, calculating genetic distances between populations from their gene frequencies and averaging them for many genes. Mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY) are transmitted by only one parent (mothers and fathers, respectively), are haploid and do not undergo recombination. Their evolution by genetic drift is therefore four times more rapid than that of autosomal genes, and they have proved particularly informative about the time range that is most useful for studies of modern human evolution — the last 100,000–150,000 years.

Population trees versus haplotype genealogies

There are two general methods for the evolutionary analysis of genetic variation. Population tree analysis usually starts by calculating genetic distances between pairs of populations from the differences in frequencies of genetic variants, averaged over many genes. Trees describe the history of population splits, after which populations diverge (it is assumed) independently. However, migration reduces distances between populations, especially those that are geographically close, generating distortions of the trees²⁴.

The other method uses genealogies of haplotypes. They are reconstructed using information from individuals rather than populations, and the branchpoints of the trees that are generated correspond to specific mutations. These genealogies are not affected by migration, and those that are based on patterns of less mutable SNPs and on insertion or deletion mutations are particularly rigorous. The dates of branching points can be calculated on the basis of knowledge of mutation rates, assuming an absence of selection. Microsatellites have much higher mutation rates and are more useful for evaluating more recent branching dates.

Principal-components analysis

This is an alternative approach that is especially useful for graphical purposes. PRINCIPAL COMPONENTS (PCs) are variables that summarize information about independent patterns that are present in a data matrix, which might be difficult to detect otherwise. PCs are ranked by the amount of information each of them reveals in the data that is being analysed. The first two or three might easily summarize more than 50% of the information that is contained in hundreds of gene frequencies²⁴. When single populations are plotted as dots on a Cartesian diagram using the first two PC values as coordinates, populations are clearly clustered by genetic similarity. If reciprocal migrations have been important in establishing similarity patterns after the principal historical separations of populations, the diagram of the first two PCs resembles their geographical map. Another mode of display is a geographical map of the values of a single PC (similar to a map of altitudes or ocean depths). This is especially useful for recognizing clines (areas of gradual variation) that occur owing to important, long-lasting migratory patterns that affect all genes equally. However, clines that are due to natural selection affect only one or a few genes. PC geographical maps have therefore been used for tracing demic expansions that started at different places and times in the same general area, and their centres of origin^{24,29}.

Current status and successes so far

The establishment of the HGDP collection, the list of populations included in it, and the conditions for obtaining DNA samples were announced in April 2002 (REF. 2). Laboratories that request samples must be non-profit-making and must send results of their studies to a CEPH database that will be made available to other researchers. One microgram of DNA — sufficient for hundreds of tests — is distributed to researchers at no cost other than shipping expenses, and larger amounts are supplied for a small charge that covers costs; however, cell lines are not distributed. The current collection consists of 1,064 cell lines from 52 populations around the world (FIG. 1). By July 2004, 56 laboratories had requested and obtained the collection.

Results from these laboratories have already been rewarding. The first researchers to use the HGDP collection, Rosenberg and colleagues⁹, genotyped each of the samples represented for 377 MICROSATELLITE loci that covered all autosomes. Their analysis of the structure of human populations was published at the end of 2002 and emphasized the importance of geographical isolation in determining genetic divergence (although other types of isolation were also observable), in agreement with the hypothesis that the divergence is mostly due to chance (random genetic drift). They also confirmed that genetic differences between populations are extremely small — in fact, smaller than those suggested by previous studies (BOX 4). Other research by the same group using the HGDP collection¹⁰ validated the estimation of early divergence

times in human evolution using microsatellites, whereas previous studies using fewer microsatellites indicated that these markers were generally useful only for dating more recent evolutionary events.

In addition, other studies that have used the HGDP resource have provided information on the usefulness of X-chromosome microsatellites¹¹; HAPLOTYPE frequencies and LINKAGE DISEQUILIBRIUM (LD) in folate-metabolism pathway genes¹²; evidence of recent positive selection at the lactase gene locus¹³ and analysis of seven Y-chromosome microsatellites¹⁴. Y chromosomes are not subject to recombination and therefore provide more information than other markers do about ancient evolutionary events¹⁵. Another study¹⁶ (discussed in more detail below) recently used HGDP samples to address general questions about sampling strategies and analysis of human genetic diversity.

The HGDP collection is the most complete worldwide human DNA collection that is available to not-for-profit researchers. The collection should prove to be an important resource for both human population genetics and evolutionary studies, as well as for biomedical studies, such as those described in the next section.

The HGDP and biomedical studies

It might be argued that the HGDP has no medical importance, because it provides no information on individual phenotypes — the only information provided about each sample is the population name, its geographical location in degrees of latitude and longitude, and the sex of each individual. However, this inference is wrong, as highlighted when, in 2003, the *Lancet* award for the year's best biomedical paper¹⁷ went to the first paper published with the HGDP data⁹. The HGDP collection is valuable not only for medical studies, but also for the study of other phenotypes, as described in later sections. It is true, however, that for most of these purposes, it would be important to increase the numbers of both populations and individuals that are currently represented.

One example of the potential use of the HGDP collection in biomedical research, particularly in countries where clinical surveys are not available or are difficult to carry out, is to estimate the incidence of recessive diseases, which are often unknown or underestimated, even if they are relatively frequent. If samples were taken from 50 individuals for each population, the detection of a single heterozygote for a mutant that is responsible for a recessive disease would indicate, according to the HARDY-WEINBERG EQUILIBRIUM, that the incidence

of individuals that express the disease should be about 1 in 10,000 (REF. 18). Most samples for populations in the current collection contain fewer than 50 individuals. However, predictions might be made more precise, not only by increasing the sample of phenotypically normal individuals from the population, but also by pooling information from genetically similar neighbouring populations.

Probably the most important medical application of the HGDP is the inexpensive provision of small but adequate and reliable control samples for ASSOCIATION STUDIES, which are increasingly being used to identify genetic variants that contribute to inherited diseases. These studies compare groups of unrelated patients from a defined population with an ancestrally similar control group¹⁹. Appropriate control samples are not easy to obtain, but could be provided by samples from collections such as the HGDP.

Another potential biomedical application is in examining the contributions of environmental factors to complex human disease.

Such analysis is usually done at the level of individuals. However, this can also be done at the population level and, in some cases — for example, for the influence of climatic and ecological factors and culture-specific customs — this approach can prove highly informative, as shown by its application to physical anthropology. In a recent successful study of this kind, data on cranial morphometry from several populations were compared with HGDP microsatellite data on genetic variation⁹ and with a database on climate. Significant correlations were found between specific cranial adaptations, patterns of genetic variation and climatic variables. One conclusion was that BRACHYCEPHALIZATION is the principal result of adaptation to extreme cold²⁰. The microsatellites examined in this study are unlikely to be located within the genes that are responsible for the observed correlations; however, they might be closely linked with genes that are responsible for some of the relevant phenotypes. Candidate genes responsible for specific phenotypes,

Box 2 | Cell lines as the basis of the HGDP collection

Ideally, the promotion of studies of human genetic diversity on the basis of comparative analysis of DNA sequences should ensure that the DNA is available to many researchers worldwide, without fear of exhausting the supply. DNA can be amplified in the laboratory in two ways: chemically, by the POLYMERASE CHAIN REACTION (PCR), and biologically, by growth of specific cell lines. Which of these approaches was more appropriate for the initiation of the Human Genome Diversity Project? Two main factors must be taken into consideration: accuracy and renewability.

Accuracy

Mitosis has been perfected in nature as a highly reliable method of DNA replication. There are indications that mutation rates are under the control of natural selection and might increase or decrease when necessary. In principle, after many growth cycles, mutations could also accumulate in cell lines that are grown in the laboratory, but we can minimize sources of error by keeping back-up subcultures of each cell line at low temperatures. The first cultures^{35,36} made for anthropological aims (in 1984) in central Africa and Bougainville are still in use and are part of the HGDP collection.

In the second half of the 1980s PCR was introduced. Direct methods of amplifying DNA by PCR have improved over the past several years, but in the absence of direct evidence, it is difficult to state how accurate they are in comparison with the replication of DNA in cells by mitosis. Rates of spontaneous mutation that take place during DNA replication of gamete-forming cells are low, on the order of 10^{-9} per nucleotide site a year, although this is difficult to estimate accurately. At the time that the HGDP was established, it seemed safer to continue to rely on cellular mitosis for DNA amplification, rather than risking the introduction of potentially higher error rates from *in vitro* methods. This is especially a source of concern for polymorphisms that occur in repetitive DNA — for example, microsatellite polymorphisms — which are much more prone to mutation than are SNPs.

Renewability

The advent of PCR means that studies are possible with much smaller samples of DNA. Nonetheless, practical experience shows that no matter how large the initial DNA samples taken are, they will eventually run out. As the aims of the HGDP dictate that many samples are taken from usually remote populations, returning to collect more DNA from the same individuals would be difficult and eventually impossible. With our increasing knowledge of genomic variation and the decreasing cost of DNA genotyping, it is becoming clear that it is preferable to study genome-scale datasets: for these purposes, renewable cell lines are still the best option. In the past, opinions differed on this point⁷, but it is notable that a major project such as the HapMap uses cell lines whenever possible³.

including medically important ones, can therefore be inferred and tested by more direct approaches. This will become more useful when the more detailed genetic information on microsatellites and SNPs that is expected from the HGDP collection becomes available, and with increased numbers of populations being available for study.

The HGDP and HapMap

Another potential biomedical application of the HGDP is its future interaction with the HapMap project^{3,4}, an ambitious research project that is aimed at solving problems of identifying the genetic determinants of complex diseases. Here, I briefly outline the aims of the HapMap project and discuss why its goals and those of the HGDP complement each other, with potentially important benefits.

The HapMap approach. During the last 50 years, substantial progress has been made in using LINKAGE MAPPING to identify genes responsible for inherited disorders that follow monogenic Mendelian patterns of inheritance. However, these disorders are responsible for a relatively small fraction of clinical cases. Most human diseases are believed to involve many genes, probably interacting in complex, non-additive ways, with potentially important environmental components.

For several reasons, the linkage approach has proved largely unsuccessful for determining the genetic basis of complex disease. The ideal approach would be to resequence the whole genome in patients and control cases to identify causal genetic variants. However, genome sequencing remains an expensive procedure, making this approach impractical. Full sequencing could be avoided by using SNPs as markers for disease-associated variants, as SNPs close to disease-related genes are likely to be transmitted with the disease. However, there is approximately 1 SNP every 1,000 nucleotides in the human genome, and a full study would require the testing of millions of SNPs per individual.

The HapMap plans to use knowledge about LD to increase the efficiency of SNP-based mapping. If haplotypes are sufficiently stable and of sufficient length, they could be used to reduce the number of SNPs that would need to be genotyped. For example, if haplotypes contained 20 SNPs on average, only 1 of these would be needed to tag a haplotype³ — the expected reduction in cost is on the order of the average number of SNPs per haplotype.

The HapMap project requires the identification of haplotypes and of at least one useful SNP in each of them. This project has already begun, and the first stage is nearing

Box 3 | Addressing ethical, legal and social issues in the HGDP

In the collection of the lymphoblastoid cell lines (LCLs) from worldwide populations, the Human Genome Diversity Project (HGDP) was acutely concerned with ethical, legal and social issues. Making sure that the needs of confidentiality and anonymity were properly addressed, that informed consent was obtained, that subjects were aware of the possible uses of the data and conforming with the legal needs of each country were the main concerns. All of the cell lines contributed to the collection were therefore reviewed to make sure that they had been collected in an ethical and legal manner¹. The US National Academy of Sciences National Research Council report⁷ provided general guidelines for this process. The background of each cell line that had been previously collected was reviewed to determine whether it was collected with informed consent for use in studies of human history or evolution. Only cell lines that complied with this were included in the HGDP resource. A protocol for confidentiality protection for donors of samples was also established. Both this and the vetting of cell lines were subsequently reviewed by an ethics advisory committee that was approved by the US National Institutes of Health Institute for General Medical Science. Other information was collected by the various researchers who contributed to the collection, but the only information that remained attached to each cell line concerned ethnic and geographical origin (in degrees of latitude and longitude), and sex.

completion. There are uncertainties about the usefulness of the HapMap data^{21–23}, because crossover events are not equally distributed, and the physical length of haplotypes and their SNP content vary greatly in different parts of the genome. Furthermore, populations vary in local LD and haplotype structure²³. There are also mechanisms of chromosome reshuffling other than crossing over that might interfere with the application of the method. Nevertheless, the project currently offers the best hope of tackling the problem of complex disease, other than a hypothetical, enormous and rapid decrease in sequencing costs.

The HapMap project will eventually cover slightly fewer than 100 individuals from each of three populations: one from Utah, with northern European ancestry, a Yoruba population from Nigeria and one from eastern Asia (Chinese and Japanese individuals). Haplotypes will be surveyed in these different populations, between which LD patterns might well vary. The extent of LD is known to be smaller in most African populations for evolutionary reasons — they have existed at least twice as long as other modern human populations and LD regions have had more time to decay because of the accumulation of crossovers. By contrast, all non-Africans — and a fraction of Africans — originate from a small population, probably of eastern African origin, that started expanding ~50,000 years ago and spread rapidly, first to Asia, and from there to other continents. The consequences of this demographical bottleneck are clearly visible in the genomes of modern humans^{15,24}, and explain their limited genetic variation and peculiar LD patterns.

Complementary, not competing. Both the HapMap and HGDP projects involve sampling

human genetic variation, but have different aims and little, if any, overlap at this stage. However, despite these differences, the two projects can only complement and reinforce each other.

The crossover events that have defined haplotypes occurred during the evolution of modern humans. Modern human history is characterized by early dispersions of an initially small population, and early rare crossovers are apparent in the geographical distribution patterns of haplotypes. Studies that use samples from HGDP populations to genotype pairs of SNPs that tag adjacent haplotypes will allow a description of the geographical distribution of the haplotypes that are defined in the HapMap project. This will help to determine the time and place of origin of rare crossovers that occurred early in human evolution, generating haplotypes that later spread to specific parts of the world and fundamentally affected the haplotype structure of human populations. As explained below, this knowledge could add a roughly equal amount to the power of current methods for evolutionary analysis. It could also be useful for the aims of the HapMap project, by increasing its analytical power in populations other than the three it has chosen to study.

Reconstructing human evolution. To clarify the way that HapMap data could complement those of the HGDP to understand the origins of human genetic diversity, it is important to appreciate that modern methods for reconstructing human evolution are based on two concurrent approaches, historical and geographical.

First, the historical reconstruction of the genealogy of mutations (the sequence of their occurrence) answers the ‘when’ questions, concerning the timing of evolutionary

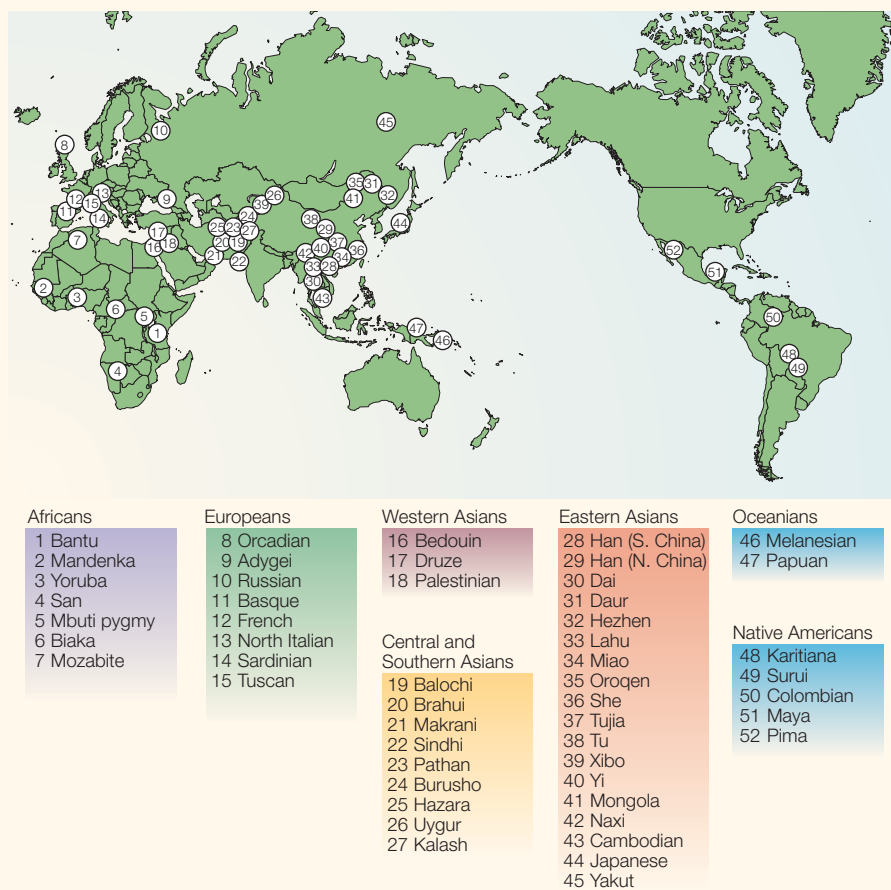


Figure 1 | **Populations that are included in the Human Genome Diversity Project collection.** A geographical distribution of the 52 populations that are represented in the Human Genome Diversity Project collection and that were used in the analysis by Rosenberg *et al.*⁹ described in BOX 4.

events. This allows the reconstruction of genealogies of extant haplotypes in the form of bifurcating trees, initiating from a common ancestor (marked by the first mutation in the species that formed modern humans)^{15,25}. These genealogies are different for every haplotype, but two haplotypes that have proved especially useful are mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY; BOX 1). They give similar genealogies¹⁵, and provide independent dates that are reasonably consistent with archaeological data. The numbers of mutations that occurred between successive splits in haplotype genealogies are used to evaluate lengths of branches in genealogies.

Knowledge of mutation rates allows an estimation of the times at which each branching occurred, including the first branching that defines the most recent common ancestor of the species for a particular haplotype. The date of the first branching is different for mtDNA and NRY (from the latest unpublished estimates, these are ~160,000 and ~100,000 years ago, with standard errors

of ~14% and ~20%, respectively). These differences are probably the consequence of variations in offspring numbers, which are greater for males, owing to polygamy.

There are several other methods of reconstructing human evolution^{24,25}. One uses distances between population pairs, calculated from the averages of large numbers of gene frequencies estimated in the populations, and is known as population tree analysis (PTA; BOX 1). Another method is based on identifying correlations between population separations and archaeological events for which the dates are known. There is good agreement between these different approaches, making the genetic reconstruction of the evolution of modern humans in the last 100,000 years fairly robust.

In contrast to this historical approach, geographical maps of the frequencies of mutations indicate 'where' they might have occurred^{26,27}. The times and places of the occurrence of mutations form the phylogeography of the species, and allow the reconstruction of the migratory paths that were taken during expansions of modern human

populations in the last 100,000 years^{15,26}.

Finally, the 'why' questions concern the processes of drift and natural selection that have affected haplotype frequencies. In the presence of natural selection, questions arise concerning the mechanism of adaptation that was involved and its identification at the molecular, biochemical, anatomical, physiological and pathological levels. The study of drift is best carried out at the whole-genome level, and that of natural selection is restricted to specific genes or interactions between genes. Both must include cultural and demographical histories of individual populations.

The HGDP facilitates these studies of the origins of human genetic diversity by providing samples of genomic DNA from populations across the world. From these, the haplotype and polymorphism data that are crucial for these studies can be obtained.

New insights from crossover events. The future availability of HapMap data will provide a potential new approach to human evolution, which could be used for reconstructing the history and geography of haplotypes by studying the times and places of the occurrence of rare, early crossovers. The evolutionary picture arising from these investigations would be complementary to that obtained from studies of the mutations that have created the current patterns of human genetic diversity. The interpretation of history is fragile because, in contrast to experimental science, no repetition of the 'experiment' is possible. But multidisciplinary approaches to the same sequence of historical events, as well as the comparison of evolutionary histories obtained from different parts of the genome and for different genetic mechanisms, provide an analogue of a repeat of the same history. They provide further opportunities to test interpretations, and can greatly strengthen our confidence in the conclusions.

These two approaches to evolution — by studying mutations or crossovers — might have similar statistical power in terms of numbers of events. We can infer from the total length of human chromosome linkage maps that more than thirty crossover events occur per meiosis, which is in the same range as the rate of SNP mutations that arise per genome per generation. It is true that the process of crossing over is likely to induce local mutations, so that the two approaches are probably not entirely uncorrelated. But it seems likely that most mutations occur independently from crossing over, and can therefore supply generally independent evolutionary evidence, although better knowledge of the relation

Box 4 | Analysis of the genetic structure of human populations using the HGDP resource

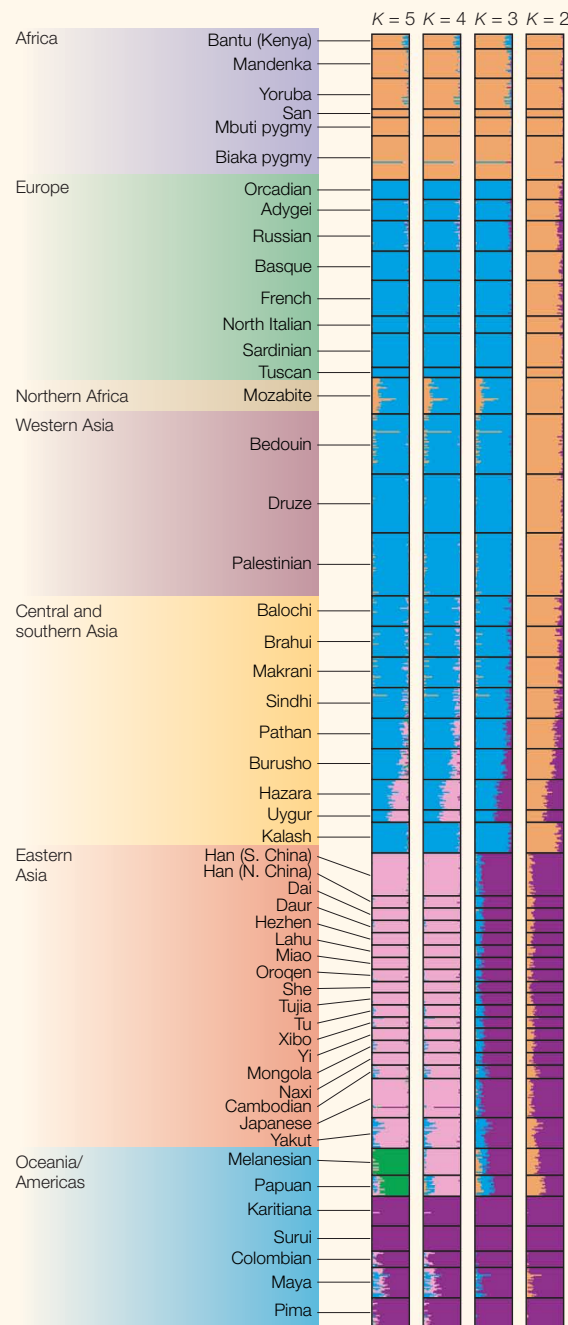
The first published study to use the Human Genome Diversity Project (HGDP) collection analysed data from 1,056 individuals from 52 populations, spanning all continents, for 377 microsatellite polymorphisms¹⁸. The study used the program Structure³⁷, which separates individuals into a number of clusters (K) that are chosen to maximize the variation between the clusters. The figure shows the results for $K = 2$ to $K = 5$, with clusters distinguished by arbitrary colours. Each vertical segment represents an individual, within which the different colours represent the proportion of admixture of the individual in terms of the clusters that represent the results. Admixture is the simplest explanation for the patterns seen, but other explanations are possible, and further analysis would be necessary to substantiate any chosen clustering pattern.

Forcing all data into 2 clusters ($K = 2$) separates all individuals and populations by longitude — with the west (Africa, Europe, and central, southern and western Asia) in one group (orange shading), and the east (eastern Asia, Oceania and the Americas) in another (purple shading). This is consistent with the beginning of the large expansion out of eastern Africa that began ~50,000 years ago and was directed predominantly eastwards. A three-cluster representation ($K = 3$) splits the western cluster of the $K = 2$ partition into sub-Saharan Africa in one group (orange shading) and Europe, western Asia and northern Africa in another (blue shading). Note that only one northern African population (Mozabites) is represented in the panel, but other results indicate that this conclusion can be extended to almost all of northern Africa. The $K = 4$ partition splits the Americas (purple shading) from eastern Asia and Oceania (pink shading), and $K = 5$ separates eastern Asia (pink shading) from Oceania (green shading).

The proportion of variation among the 52 HGDP populations estimated in REF. 9 is extremely small compared with that from within populations. Analysis of variance using F_{st} values (which are a common measure of gene frequency variation) indicates that, from these results, only 5–7% of the world genetic variation is accounted for by differences between the 52 populations⁹. Differences between the 5 principal groups account only for 3–5%, and differences between populations that are within a major group account for 2–4%. Clearly, the fraction of variation among individuals within populations — that is, the residual 93–95% — is by far the most important portion. This estimate is definitely larger than that observed previously from protein data, as well as from DNA polymorphisms, (including microsatellites that were tested previously³⁸), which was 85% (REF. 39). However, microsatellites are expected to give lower F_{st} variances than SNPs because they have lower gene frequencies, having on average many more alleles^{24,38}. Moreover, earlier estimates³⁹ were obtained with fewer (14), more heterogeneous population clusters than the 52 HGDP populations. Grouping the 52 HGDP populations data to mimic the 14 clusters used in REF. 38 decreased the estimate of the variation within populations to 89.8%.

Although the variation observed among the 52 HGDP populations with the microsatellite set used⁹ is smaller than any earlier estimate, it allows us to reconstruct¹¹ a population history that is consistent with the standard model of human evolution^{10,15}.

Reproduced, with permission, from REF. 9 © (2002) American Association for the Advancement of Science.



between the two processes would be beneficial. Using the LD approach, differences between populations or individuals based on LD comparisons would replace those that are based on population allele frequency or individual SNP difference. So, with increasing knowledge of HapMap tags, crossover history and geography might provide useful additional information to the evolutionary history of the human species that can be investigated using data from projects such as the HGDP.

Implications for the HapMap. An increased knowledge of LD patterns might also allow the application of the HapMap approach to medical problems within populations other than the three now being studied. LD blocks in some parts of the genome are likely to be useful in some parts of the world and not in others, because early crossovers might have destroyed LD of some haplotypes in some regions of the world and not in others⁹. It would therefore

seem that cooperation between the HGDP and the HapMap is highly desirable, even at this early stage.

The future of the HGDP

The main limitations of the current HGDP collection are that the present list of populations is small and does not evenly cover the inhabited regions of the world. In addition, the number of individual samples (1,064) is small, although this has already

allowed studies that have provided some interesting conclusions, as described above.

The main future requirement for the project is clearly to increase the number of cell lines, especially from areas that are now insufficiently represented. There is currently a greater concentration of samples from countries that have pioneered the idea of collecting cell lines from different ethnic groups and making them available for research — such as Israel, Pakistan and China. By contrast, India and Polynesia are not represented at all, and Europe, northern Asia, the Americas and Oceania have limited representation. Population samples in the collection contain an average of 20 individuals — about the size that was decided on as a compromise between suggestions from the first HGDP symposium. However, the sample size per population varies from 1 to 50 individuals. Populations from Pakistan and China represent various ethnic groups, and their samples are of 10 individuals, which is small by most criteria. One solution is to pool data from neighbouring populations that are sufficiently genetically similar.

Another question is whether the HGDP should focus in the future on individuals as the unit of sampling, or whether the emphasis should remain on sampling populations. These alternatives were considered at the start of the project. This is worth considering again now because of a recent paper¹⁶ emphasizing that, for interpreting human genome diversity,

attention to clines (gradual variations of populations in space) is preferable to focusing on clades (which results from a history of sharp population separations). Clines are certainly common for many genes²⁴, which is one reason to criticize the use of the distinction of races in humans, as already emphasized by Charles Darwin. Different methods of analysis emphasize either one or the other interpretation. The construction of trees by PTA (REF. 28) forces population data into clades, whereas multivariate principal component analysis (BOX 1), especially if displayed as geographical maps^{24,29}, tends to turn them into clines. Which interpretation is more accurate? This depends on local history and genetic geography, and understanding these factors should be an important aim of any study. There are geographical, linguistic and social barriers between populations, and the history of population separations — when these are sharp and not greatly altered by later migration — tends to generate discontinuities. When these discontinuities are real, population trees might be meaningful. DEMIC EXPANSIONS create strong clines for many genes, with clear centres of origin, but their multiplicity in a particular area (probably a common occurrence) generates a local mixture of clades and clines (BOX 1). Natural selection often creates clines of individual genes; individual migration tends to create clines for all genes, but group migrations to remote areas create new clades³⁰.

Different methods of analysis might emphasize an interpretation in terms of clines or clades, but the important question with respect to the HGDP is whether the sampling method used so far has irreversibly affected the analysis of data provided by the collection. Undoubtedly, data collection that involves gross or fine clustering of individuals into populations will strongly affect the perception of clines or clades¹⁶. Therefore, should the choice of the sampling unit for the HGDP be changed in the future from populations to individuals collected at uniform distances, as in one original suggestion? It is logistically more efficient to collect population samples, and it is certainly better than collecting random individuals at specified distances. It is also necessary, especially for the study of recessive alleles, to test whether random mating conditions and the absence of natural selection are satisfied, which might be difficult to do with individually based collections. Ignoring the social realities of populations also seems dangerous. For example, a naive sampling that is based on geographical distances between individuals in New York city would only generate a badly biased history of the whole world. Because of the narrow geographical range of most migration at the level of individuals, the similarity of geographically close populations is so strong^{15,24} that it seems reasonable to continue sampling small, well-defined populations of obvious anthropological or medical interest. Therefore, it seems reasonable to proceed in the direction followed so far by the HGDP. It is also comforting to notice that the sizes of populations sampled by the HGDP are small enough that there has been a substantial reduction in inter-population variance compared with all estimates from earlier population collections (BOX 4).

The danger of forcing cladistic interpretations¹⁶ using HGDP data seems remote, especially when the collection will be increased sufficiently in the future to remove major discontinuities in the present geographical distribution of populations. Moreover, the tendency of humans to cluster into social groups has important social and medical implications that might be lost if sampling was carried out at regular geographical intervals, rather than from small social groups. However, it is certainly interesting to test various random sampling schemes, and this might be made possible when national DNA collections become available.

Bearing in mind these considerations, strategies for extending the HGDP include asking teams of scientists that are collecting blood samples for anthropological or medical purposes to donate a fraction of their blood

Glossary

ADMIXTURE

The mixture of two or more genetically distinct populations.

ASSOCIATION STUDIES

A method for localizing genes that are responsible for specific diseases by comparing the DNA of a selected set of patients who are believed to carry the same mutation/s because of their ancestral origin, with that of unrelated healthy controls from the same population.

BRACHYCEPHALIZATION

An increase in the breadth to length ratio of the skull.

DEMIC EXPANSIONS

Processes of substantial demographical growth causing geographical expansions of a population. These are made possible by innovations that affect production of food, such as agro-pastoral economies and/or other improved technologies (for example, transportation, hunting and other weapons).

HAPLOTYPE

A set of genetic markers that show complete or nearly complete linkage disequilibrium; that is, they are inherited through generations without being changed by crossing-over or other recombination mechanisms.

HARDY-WEINBERG EQUILIBRIUM

A classical mathematical principle in population genetics used for testing random mating. It gives the expected frequencies of genotypes for a gene after one generation of random mating if the parental allele frequencies are known.

LINKAGE DISEQUILIBRIUM

The tendency for markers that are physically close to each other on the same chromosome to be transmitted to the progeny together, as there is a low probability that they will be split through recombination.

LINKAGE MAPPING

Mapping genes by typing genetic markers in families to identify regions that are associated with disease or trait values that occur within pedigrees more often than is expected by chance. Such linked regions are more likely to contain a causal genetic variant.

LYMPHOBLASTOID CELL LINES

Lymphoblastoid cell lines are obtained from B lymphocytes, a fraction of white cells from blood that can be grown indefinitely in the laboratory after special treatment of the cells with Epstein–Barr virus.

MICROSATELLITES

Microsatellites are tandem repeats of short nucleotide sequences (2–6 bases). They have a large number of alleles compared with SNPs, owing to a much higher mutation rate.

samples for making cell lines, continuing the tradition that built the HGDP collection. Sample donations from countries that are likely to generate national collections might allow an inexpensive enlargement of the present HGDP collection. National DNA and cell-line collections, known as biobanks, are being planned or established in Canada, Estonia, Iceland, Italy, Norway and the United Kingdom, and more are likely to follow. These biobanks could donate LCLs (or blood samples from which LCLs could be made), representing a minute fraction of their collection, ideally forming a geographically (and if necessary, linguistically and ethnically) stratified sample of the country. The only sample currently included in the HGDP that corresponds to this description is from France, and was generated by the CEPH at my suggestion.

From an ethical point of view, studies of human population genetics and evolution have generated the strongest proof that there is no scientific basis for racism, with the demonstration that human genetic diversity between populations is small, and perhaps entirely the result of climatic adaptation and random drift^{9,15,24}. It is to be hoped that the fears that were associated with the analysis of human variation have largely disappeared, and expansion of the study of human genetic diversity can become more efficient and scientifically satisfactory. The HGDP can help to achieve this aim. However, this project has survived with little support until now, and will need an increased level of funding. Its potential uses in medicine, science and social problems such as racism are sufficiently important that the project should be continued and expanded.

L. Luca Cavalli-Sforza is at the Genetics Department, Stanford Medical School, Stanford, California 94305, USA.

e-mail: cavalli@stanford.edu
doi:10.1038/nrg1579

1. Greely, H. T. Human genome diversity: what about the other human genome project? *Nature Rev. Genet.* **2**, 222–227 (2001).
2. Cann, H. M. *et al.* A human genome diversity cell line panel. (letter) *Science* **296**, 261 (2002).
3. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–795 (2003).
4. The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nature Rev. Genet.* **5**, 467–475 (2004).
5. Cavalli-Sforza, L. L. How can one study individual variation for three billion nucleotides of the human genome? *Am. J. Hum. Genet.* **46**, 649–651 (1990).
6. Cavalli-Sforza, L. L., Wilson, A. C., Cantor, C. R., Cook-Deegan, R. M. & King, M.-C. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics* **11**, 490–491 (1991).
7. Committee on Human Genome Diversity, National Research Council. *Evaluating Human Genetic Diversity* (US National Academy of Sciences, Washington DC, 1997).

8. Dausset, J. *et al.* Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577 (1990) (in French).
9. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
10. Zhivotovskiy, L. A., Rosenberg, N. A. & Feldman, M. W. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72**, 1171–1186 (2003).
11. Ramachandran, S., Rosenberg, N. A., Zhivotovskiy, L. A. & Feldman, M. W. On the robustness of the inference of human population structure. *Hum. Genomics* **1**, 87–97 (2004).
12. Shi, M., Caprau, D., Romitti, P., Christensen, K. & Murray, J. C. Genotype frequencies and linkage disequilibrium in the CEPH Human Diversity Panel for folate pathway genes *MTHFR*, *MTHFD*, *MTFR*, *RFLJ* and *GCP2*. *Birth Defects Res. A* **67**, 545–549 (2003).
13. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
14. Macpherson, M. J., Ramachandran, S., Diamond, L. & Feldman, M. W. Demographic estimates from Y-chromosome microsatellite polymorphisms: analysis of a worldwide sample. *Hum. Genomics* **1**, 345–354 (2004).
15. Cavalli-Sforza, L. L. & Feldman, M. W. *Biology as history: population genetic approaches to modern human evolution.* *Nature Genet.* **33**, 266–275 (2003).
16. Serre, D. & Paabo, S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685 (2004).
17. Horten, R. *et al.* Read all about it: the *Lancet's* paper of the Year, 2003. *Lancet* **362**, 2101–2103 (2003).
18. Cavalli-Sforza, L. L. & Bodmer, W. *The Genetics of Human Populations* (Freeman, San Francisco, 1971; Dover, New York, 1999).
19. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
20. Roseman, C. C. Detecting inter-regionally diversifying natural selection of modern human cranial form using matched molecular and morphometric data. *Proc. Natl Acad. Sci.* **101**, 12824–12829 (2004).
21. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
22. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
23. McVean, G. A. T. *et al.* The fine scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
24. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, New Jersey, 1994).
25. Cavalli-Sforza, L. L. The DNA revolution in population genetics. *Trends Genet.* **14**, 60–65 (1998).
26. Underhill, P. A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62 (2001).
27. Edmonds, C. A., Lillie, A. S. & Cavalli-Sforza, L. L. Mutations arising in the wave front of an expanding population. *Proc. Natl Acad. Sci. USA* **101**, 975–979 (2004).
28. Cavalli-Sforza, L. L. & Edwards, A. W. F. Analysis of human evolution. *Genet. Today Proc. 11th Int. Congress Genet.* **3**, 923–933 (1964).
29. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. L. Synthetic gene frequency maps in Europe. *Science* **201**, 786–792 (1978).
30. Cavalli-Sforza, L. L. Some current problems in human population genetics. *Am. J. Hum. Genet.* **25**, 82–104 (1973).
31. Hirsfeld, L. & Hirsfeld, H. Essai d'application des methodes au probleme des races. *Anthropologie* **29**, 505–537 (1919) (in French).
32. Race, R. R. & Sanger, R. *Blood Groups in Man* (Blackwell Scientific, Oxford, 1975).
33. Pauling, L., Itano, A. H., Singer, S. J. & Wells, I. C. Sickle cell anemia, a molecular disease. *Science* **110**, 543–548 (1949).
34. Harris, H. *The Principles of Human Biochemical Genetics* 3rd edn (Elsevier; North Holland Biomedical Press, Amsterdam, 1980).
35. Cavalli-Sforza, L. L. *et al.* DNA markers and genetic variation in the human species. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 411–417 (1987).
36. Cavalli-Sforza, L. L. (ed.) *African Pygmies* (Academic, Orlando, 1986).
37. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
38. Bowcock, A. M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).
39. Barbujani, G. *et al.* An apportionment of human DNA diversity. *Proc. Natl Acad. Sci. USA* **84**, 4516–4519 (1987).

Acknowledgements

This work has been made possible by donors of blood samples and cell lines to the Human Genome Diversity Project (HGDP) and the Center for the Study of Human Polymorphism (CEPH). The collaboration with CEPH has been a decisive contribution. Support for preparing the first African cell lines in the Stanford laboratory in 1984–1985 came initially from the Lucille P. Markey Trust, with later additions from a National Institutes of Health Institute for General Medical Science programme and the HGDP–CEPH initiative from the Ellison Medical Foundation. H. Cann, M. Feldman, H. Greely and M.-C. King are thanked for suggesting improvements to the manuscript.

Competing interests statement

The author declares no competing financial interests.

 **Online links**

FURTHER INFORMATION

- Fondation Jean Dausset — CEPH:**
http://www.cephb.fr/ceph_presentation.html
- International HapMap Project:** <http://www.hapmap.org>
- Human Genome Diversity Project:**
<http://www.stanford.edu/group/morrinst/hgdp.html>
- Human Genome Organisation:**
<http://www.gene.ucl.ac.uk/hugo>
- Human Genome Project:**
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
- Marcus Feldman's laboratory:**
<http://charles.stanford.edu/pubs.html>
- National Research Council:**
<http://www.norveco.com/html/lab/NAS-NRC.htm>
- Noah Rosenberg's web site:**
<http://www.cmb.usc.edu/people/noahr/projects.html>
- Stanford Human Population Genetics Laboratory:**
<http://hppl.stanford.edu>
- Access to this interactive links box is free online.**

ERRATUM

EMERGING TECHNOLOGIES FOR GENE MANIPULATION IN *DROSOPHILA MELANOGASTER*

Koen J. T. Venken and Hugo J. Bellen

Nature Reviews Genetics **6**, 167–178 (2005); doi:10.1038/nrg1553

In this article the Cre recombinase was incorrectly defined as a cyclic AMP-response element. This correction has been made to the online enhanced text and PDF version of this review.