DOE/ER/61463--T3

CONF-9309530--Summ.

The Human Genome Diversity (HGD) Project

SUMMARY DOCUMENT

incorporating
the HGD Project outline and development,
proposed guidelines,
and
report of the International Planning Workshop
held in Porto Conte, Sardinia (Italy)
9-12th September 1993

Financial support for the workshop was provided by the Porto Conte research and Training Laboratories Foundation, Sardinia; the European Commission; the Soros Foundation; the United States National Science Foundation, National Institutes of Health and Department of Energy; HUGO Europe.

This report has been compiled on behalf of the Human Genome Diversity (HGD) Committee of HUGO, the Executive Committee for the global HGD Project:

Dr Julia Bodmer (UK)
Dr Walter Bodmer (UK)
Dr Luca Cavalli-Sforza (USA) - Chairman
Dr Marc Feldman (USA)
Dr Takashi Godjobori (Japan)
Dr Ken Kidd (USA)
Dr Mary-Claire King (USA)
Dr Partha Majumder (India)
Dr Onesmo ole-MoiYoi (Kenya)
Dr Alberto Piazza (Italy)
Dr Svante Pääbo (Germany)
Dr Marcello Siniscalco (Italy)
Dr Ken Weiss (USA)

Dr Liz Evans
Secretary, HGD Executive Committee

HUGO Europe, One Park Square West, London NW1 4LJ, UK
Tel: [44] (71) 935 8085 Fax: [44] (71) 935 8341

PREFACE

In 1991 a group of human geneticists and molecular biologists proposed to the scientific community that a world wide survey be undertaken of variation in the human genome. The Human Genome Organisation (HUGO) responded to this proposal by establishing an *ad hoc* committee to consider how such a global project could be developed. The likely complexity of the project soon became clear as it was seen to encompass the interests of biomolecular scientists, human geneticists, anthropologists, archaeologists, evolutionists, linguists and historians. To aid their considerations, the committee therefore decided to hold a small series of international workshops to explore the major scientific issues involved. The intention was to define a framework for the project which could provide a basis for much wider and more detailed discussion and planning - it was recognised that the successful implementation of the proposed project, which has come to be known as the Human Genome Diversity (HGD) Project, would not only involve scientists but also various national and international non-scientific groups all of which should contribute to the project's development. The international HGD workshop held in Sardinia in September 1993 was the last in the initial series of planning workshops. As such it not only explored new ground but also pulled together into a more coherent form much of the formal and informal discussion that had taken place in the preceeding two years. This report presents the deliberations of the Sardinia workshop within a consideration of the overall development of the HGD Project to date. Arising from the discussions at the workshop was the specific request that the continuing development of the HGD Project be overseen by HUGO. A formal proposal to this effect was put to and approved by the Council of HUGO in January 1994.

Contents

Contents (contd.) Page(contd.)

I THE HUMAN GENOME DIVERSITY (HGD) PROJECT

Aims

The Human Genome Diversity Project is a collaborative research project that is being developed on a global basis under the auspices of the Human Genome Organisation (HUGO). The overall goal of the project is to arrive at a much more precise definition of the origins of different world populations by integrating genetic knowledge, derived by applying the new techniques for studying genes, with knowledge of history, anthropology and language. More specifically the aims are:

To investigate the variation occurring in the human genome by studying samples collected from populations that are representative of all of the world's peoples, and

Ultimately, to create a resource for the benefit of all humanity and for the scientific community worldwide. The resource will exist as a collection of biological samples that represents the genetic variation in human populations worldwide and also as an open, long-term, genetic and statistical database on variation in the human species that will accumulate as the biological samples are studied by scientists from around the world.

Value

1 The main value of the HGD Project lies in its enormous potential for illuminating our understanding of human history and identity.

2 The resource created by the HGD Project will also provide valuable information on the role played by genetic factors in the predisposition or resistance to disease.

3 The HGD Project will bring together people from many countries and disciplines. The work of geneticists will be linked in an unprecedented way with that of anthropologists, archaeologists, biologists, linguists and historians, creating a unique bridge between science and the humanities.

4 By leading to a greater understanding of the nature of differences between individuals and between human populations, the HGD Project will help to combat the widespread popular fear and ignorance of human genetics and will make a significant contribution to the elimination of racism

II INTRODUCTION TO THE HUMAN GENOME DIVERSITY (HGD) PROJECT

The cells of everybody alive today, regardless of where or how they live, contain the same 100,000 or so genes. Collectively known as 'the human genome', these genes contain all the information that makes us appear and function as humans rather than as members of some other species. Identifying each of the genes and locating its position on one of the human chromosomes is the first aim of the Human Genome Project, the on-going research project to which scientists from many countries are contributing. However, many human genes exist in more than one form (or 'allele') and we do not all carry exactly the same forms of every variable ('polymorphic') gene. Each of us, apart from identical twins, is thus a unique individual, recognisably human but different from all other humans. The genetic variation from one person to another reflects the evolution of our species as it is the result, over many generations, of the survival or loss of different forms of genes or the natural introduction of new forms. Studying this variation among people from around the world, which is the aim of the newly-developing Human Genome Diversity (HGD) Project, can therefore provide a great deal of information about the development of our species which, integrated with findings from archaeology, linguistics, history and other disciplines, can lead to a much richer and more complete picture of our past than has previously been possible.

It is expected that the results of the HGD Project will lead to a much greater understanding of the history of modern populations and their precursors - where people came from, what geographical routes took them to their present territories, what types of technical innovations they were responsible for, how they interacted socially within their populations and with other populations over the course of history, why their different traits or languages may have developed, whether there were major reductions in the size of populations at different times due to catastrophes such as infectious diseases, and many other issues. As a cultural resource the potential of the project is therefore enormous.

The HGD Project will also provide the scientific data to confirm and support what is already clear from populations studies - that, in biological terms, there is no such thing as a clearly defined race. Biologically, there is a continual graduation from one population to another: populations are defined on a statistical basis rather than on the basis that each has entirely distinct and different genetic or physical characteristics. Most importantly, therefore, the results of the Project are expected to undermine the popular belief that there are clearly defined races, to contribute to the elimination of racism and to make a major contribution to the understanding of the nature of differences between individuals and between human populations.

Studying genome variation in different populations will also provide valuable information about the factors contributing to disease. The incidence of different diseases varies between populations and, while much of this variation can be explained by environmental factors such as diet, climate, parasites, infectious diseases or a whole range of 'pollutants', genetic factors are also known to have a predisposing effect in many cases. Identifying these factors (as well as those genetic factors that are responsible for resistance to disease) and studying their distribution in different populations will increase the likelihood of being able to

develop more effective ways of treating or preventing many diseases and may have direct practical consequences in terms of the provision of health care resources. The information arising from the Project will also be critically important for understanding basic biological phenomena such as the inheritance of disease, the development of cancer and the processes of ageing. It will also provide base-line information about human genome variation that will be useful in identifying individual samples for paternity, forensic and other applications.

Using language as a criterion, there are over 5,000 distinct human populations in the world and it would clearly be beyond the means of any global project to study them all. For the Human Genome Diversity Project it is initially proposed to study genome variation in several hundred populations which - in ways discussed later in this document - can be considered to be broadly representative of all. To achieve this, the selected populations must be truly representative of a given area and defined by carefully considered anthropological criteria.

Samples from individuals within each of these populations will be collected and the DNA content analysed to produce data on the frequency of occurrence within the population of an agreed set of alleles or other genetic markers. In order to be able to define the relationships between populations, a set of core markers will be studied in all the populations. In addition to the core markers, a wide variety of other markers will eventually be studied, including those that have clinical relevance or a particular relevance to regional subsets of populations. The total number of samples per population on which these markers are tested will need to be large enough to ensure that the population can be characterised by a distribution of marker frequencies and must not be too biased with respect to location - several 'villages' within an area would need to be sampled, for example. A pilot project is being developed to explore the distribution of a wide range of markers and provide a basis on which the initial choice of core markers can be made.

In order to establish a resource that will be available for many years and that will allow future scientists to study any polymorphism that is defined at the DNA level, the intention is to create an unlimited supply of DNA from each of the collected samples. For most samples (of blood, hair or cheek scrapings) DNA will be extracted from the sample and stored for long-term use but, in order to provide a back-up source of 'original sequence' DNA, a number of the blood samples will be used to develop cell lines. In principle, a finite amount of stored DNA can be made to last indefinitely through the application of PCR (polymerase chain reaction)-based techniques, which enable multiple copies to be made of even very small amounts of DNA.

Genetic population studies are not new. They have been carried out for most of this century, mainly by researchers in the biomedical sciences, by anthropologists and by linguists. What is new is the possibility of extending the study of population to a much more detailed level by applying some of the DNA technology (such as the PCR-based technology mentioned above) that has been developed within the last few years in the context of the Human Genome Project. The availability of this technology has revolutionised the study of the human genome and has already revealed a level of variation between individuals which is far greater than anything previously detected. As a result, the precision with which populations, their origins and their interrelationships can be defined, using relatively small samples, has increased enormously. What is also new is the concept

of using a common set of markers with all the population samples that are studied, so providing a systematic basis for comparing the data from different populations.

The definition of a polymorphism, together with the number of samples that it is feasible to consider handling for each population, may place constraints on the minimal frequency of variants that will be detected reliably in the samples collected for the HGD Project. For this reason, wherever possible, samples should be as large as practical.

Ironically, at the same time as advances in technology have made it possible to undertake a detailed study of human genome variation, the human species is moving towards increasingly intensive amalgamation. Human populations have probably always been in flux but there is widespread interest in being able to reconstruct the dynamics of human populations in the time prior to known or written history ('prehistory'), particularly in the time before the dislocations caused by the large-scale transoceanic/continental migrations of recent millenia. This leads to an interest in sampling those of the 'native' or 'aboriginal' populations in each region - descendants of peoples present at the time of major incursions from other continents - who seem likely to have been least affected by admixture with the incoming populations. Study of these populations optimizes the ability to reconstruct the ethnographic map to its state at the beginning of recorded history. Such ethnographically based data will improve the ability to understand the population dynamics of each region prior to this. In the absence of such sampling, the rate at which admixture and population amalgamation are taking place today is so great that in a few generations much of the valuable information about regional prehistory will be made very difficult to reconstruct.

The HGD Project, however, is not confined to an interest in the 'native' peoples of each region or to the reconstruction of population history. Most people on earth today are members of very large, cosmopolitan populations formed by complex patterns of major recent admixture, largely due to rapid long-distance transportation. It is important to understand the variation, historical dynamics and demography of these populations. In particular, genetic data can be important in the effective application of advances in molecular medicine to all of the world's peoples. Thus the HGD Project is designed to represent all of our highly variable species, past and present. The Project was proposed to take advantage of the information that can be gleaned for that purpose by the powerful new techniques for studying DNA.

Many of the population studies carried out in the past were done under widely differing conditions. Consequently, while an enormous amount of important and useful information has been obtained from individual studies, it has frequently proved difficult to compare results from one investigation to another. Information and samples are also dispersed in laboratories around the world and access to them is often limited. To overcome such problems, as well as to achieve its aims in the most efficient, cost-effective way, the Human Genome Diversity Project is being planned from the outset on a global basis with world-wide coordination of sampling and testing. Local participation in all areas of the world will be essential and the success of the Project will also be entirely dependent on international collaboration and cooperation. To promote global involvement and to coordinate activities internationally, the HGD Project is being developed under the auspices of the

Human Genome Organisation, HUGO, the international organisation of scientists already coordinating activities within the Human Genome Project.

Like many scientific endeavours, the Human Genome Diversity Project has ethical as well as practical aspects. In contrast to some other endeavours, however, the importance of this ethical dimension has been recognised from the outset and has already been extensively discussed as an integral part of the HGD Project's planning process. The most sensitive issues are seen to concern the protection of sampled populations and individuals and the preservation of their rights in the sampled materials. These issues are considered in this document together with all the major scientific aspects of the HGD Project that were discussed at the workshop in Sardinia.

III THE CASE FOR THE HGD PROJECT

A scientific contribution to world culture

At present, differences in the distribution of gene (or marker) frequencies provide the only scientifically-objective basis for defining human populations from a biological point of view. From this basis, well established theories and statistical procedures have been developed for analysing relationships between populations (including the construction of phylogenetic trees) and for determining patterns of migration and population admixture. These approaches provide the only biologically valid way to trace much of human history and the recent evolution of *Homo sapiens* and there are many examples of their application using classical blood groups, HLA and protein electrophoretic polymorphisms. However, such examples give only a small indication of the range of variation that can be studied with the DNA markers that are now available or of the extent of variation that it will be possible to study in the future as yet more markers become available and as laboratory procedures become automated.

Studies on the distribution of DNA marker frequencies in different populations have already thrown some light on a number of major population issues. For example, we have gained a much clearer idea of the young age of our species, strengthening the evidence that we had our origin in Africa at some point about 100,000 years ago. For most continental regions, we have genetic evidence for major demographic expansions in the last 10,000 to 50,000 years. Patterns of genetic variation among and between populations speaking the major language groups in th the Americas are being used to develop models of the number and timing of the major immigration events that occurred across the Bering Straits. Settlement patterns of the island Pacific, including Papua New Guinea and Australia are being reconstructed and the origins of those populations from different regions of Southeast Asia are being identified. Detailed studies within Europe have shown genetic evidence for major historical migration events, as well as providing evidence for the demographic nature of the spread of farming technology into Europe around 10,000 years ago.

These are just examples of the types of analyses whose power will be greatly enhanced by the enormous increase in the number of polymorphisms that can now be studied at the DNA level, a number which will grow exponentially in coming years. It is expected that we will develop a much sharper understanding of the history and prehistory of many groups of long-standing interest to human sciences, such as the Ainu in Japan, the Lapps and Basque populations of Europe, the Hadza in Africa who live in the north but speak a language otherwise found only in the south, the different populations of the arctic, and many other examples. Good genetic data will enable us to develop a better understanding of the biological history of social subdivisions such as the caste system of India and, where intensively sampled, of the local social microdemography by which most of the dispersion of human variation occurs. And we will gain an understanding of the key role India has played in the human history and settlement of Eurasia. In addition, as techniques develop for obtaining DNA from samples of bone that may be thousands of years old, it should be possible to obtain better time depth in our studies of human population history. For example, we may be able to gain a greater

understanding of population history prior to the disruptions caused by the age of global European expansion.

The primary case, therefore, for the Human Genome Diversity Project is cultural. It will also provide information on the evolution of the human genome, the geographic distribution of human variation, the age of populations and evidence of major migrations, interesting patterns of migration and other phenomena of that nature. The study of genetic polymorphism in human populations creates a unique bridge between the science of human genetics and the humanities, including anthropology, archaeology, history and linguistics, and presents scientists with a unique opportunity to contribute to the world's cultural heritage. There is a cultural imperative for us to respond to that opportunity and use the extraordinary scientific power that has been created through the development of DNA technology to generate - for the benefit of all people - information about the history and evolution of our own species.

An essential basis for genetic epidemiology

In addition to its cultural importance, the data generated by the HGD Project will have important practical uses. The most far reaching of these will be in the application of the data to the study of disease. Genetic epidemiology is now believed by many to be the epidemiology of the future but it cannot be developed without a sound basis of knowledge of genetic marker distributions in properly sampled normal populations. Every time we ask whether a particular genetic marker is asociated with a disease, we need to know about the normal control populations. The need for this comparison increases with the diversity of the populations. The population of the United States, being extraordinarily diverse, presents an outstanding example of this epidemiological need. What is measured or observed within that populations needs to be defined in relation to population samples throughout the world - not just Europe but also Africa and all of Asia, including India, China, Japan, the Pacific and the Americas.

There are other specific examples of the value of the genetic typing of populations for disease studies. For instance, it was population studies that were largely responsible for revealing the HLA system and, over the last thirty years, studying the distribution of HLA polymorphisms in normal populations has been intrinsic to the study of the HLA system and of its functional significance. We have learned of the need to know the distribution of HLA types for transplant matching studies. We can also achieve a good indication of the distribution of HLA associated diseases in different populations - the extremely high association between HLA-B27 and ankylosing spondylitis, for instance, essentially determines the incidence of ankylosing spondylitis within a population. Planning vaccines based on a knowledge of peptide motifs that associate with different HLA alleles also requires a knowledge of the HLA distribution within a population, which is likely to be a major determinant in the design of peptide-based vaccines.

Linkage disequilibrium - the extent of association within a population of two very closely linked genetic polymorphisms - can be a major clue to positional cloning when a genetic variant is too common to be maintained by a mutation-selection balance. A significant example of this is the cloning of the cystic fibrosis gene. In order to use a marker for linkage disequilibrium studies, it is essential to know its

distribution in different normal populations. It is only in this way that it is possible to establish, to some extent at least, the age of a variant and therefore its usefulness for linkage disequilibrium studies. A major study has recently been undertaken on the origin of the CF 508 deletion in European populations, by studying associated micro-satellite haplotypes. The interpretation of this data, from which the age of the deletion can be estimated, depends on having population samples from which the nomal distribution of the microsatellite-based haplotypes (ie. not associated with the CF 508 mutation) can be established.

Every study of the association of genetic markers with disease requires data on a properly chosen control population. A number of studies have already produced very valuable results and it is undoubtedly on this basis that studies of the role of genes in multifactorial diseases (such as Alzheimer's) will be carried out in future. Recent examples include the association of a P450 variant with Parkinson's disease, of ACE variants with heart disease, of non-HLA associated variations with diabetes and the intriguing indication of an asthma-associated variation in an IgE receptor. Future studies will include exploring the association of repair deficiency heterozygotes with cancer, and investigating the presence of low penetrance, high frequency alleles of tumour suppressor genes by looking for associations between cancer and polymorphisms within the genes.

IV INTERNATIONAL PLANNING WORKSHOP FOR THE HGD PROJECT 9-13 September 1993, Sardinia, Italy

Background

In 1991, Luca Cavalli-Sforza, together with the late Allan Wilson, Charles Cantor, Bob Cook-Deegan and Mary-Claire King, published a paper in *Genomics* (Vol 11, pp.490-91) proposing that a world-wide and geographically comprehensive survey of human genome variation be undertaken as quickly as possible. The aim was to apply the powerful tools of molecular genetic analysis to a systematic study of human populations. These would be chosen using agreed scientific criteria and would include, in part, certain populations that have so far been relatively isolated but which may not remain so for much longer.

Walter Bodmer - a population geneticist and, at the time, President of the Human Genome Organisation (HUGO) - responded to the proposal by suggesting that Luca Cavalli-Sforza request the setting up of an *ad hoc* committee under HUGO to explore the development of a Human Genome Diversity (HGD) project. This led to an informal discussion among a group of interested individuals attending the International Congress of Human Genetics in Washington in late 1991, including representatives from the United States National Institutes of Health, the National Science Foundation and the Department of Energy, at which it was suggested that three planning workshops should be arranged to prepare the scientific case for the project. The workshops would involve molecular and population geneticists, anthropologists, linguists and archaeologists and would be concerned with issues such as sampling procedures, the choice of populations, the techniques to be used and the markers to be studied. The Council of HUGO approved the setting up of the *ad hoc* HGD Committee under the Chairmanship of Luca Cavalli-Sforza, and nominated Walter Bodmer to liaise with the committee on their behalf. The membership of the committee was as follows:

Dr Julia Bodmer (UK)
Dr Walter Bodmer (UK)
Dr Luca Cavalli-Sforza (USA)
Dr Marc Feldman (USA)
Dr Ken Kidd (USA)
Dr Mary-Claire King (USA)
Dr Svante Pääbo (Germany)
Dr Alberto Piazza (Italy)
Dr Marcello Siniscalco (Italy)
Dr Ken Weiss (USA)

The first two scientific planning workshops for the HGD Project were held in the United States in 1992 with the help and support of the National Science Foundation, the National Institute of General Medical Sciences, the National Center for Human Genome Research, the National Institutes of Health and the Department of Energy. The first workshop, organised by Luca Cavalli-Sforza and Marc Feldman took place at Stanford University in July 1992 and concerned the statistical issues of sampling populations. The topics discussed included sampling and population structure, analysis of populations, drift versus natural selection, modelling migration and population subdivision, and population structure and subdivision. The second workshop, organised by Ken Weiss, took place at

Pennsylvania State University in November 1992 and focussed exclusively on anthropological issues. The aim of this workshop was to identify those issues most pertinent to the selection of representative populations from each area of the world and to begin to propose examples of particular populations that would meet the defined criteria. An additional meeting was held at the end of 1992 at the National Institutes of Health in Washington (and at their behest) to discuss technical and ethical issues that would arise within the HGD Project.

During 1992 discussions of Human Genome Diversity also took place in several other contexts under the auspices of HUGO. These included the *Human Genome 92* meeting in Nice (France), the second HUGO Europe meeting in Sardinia (Italy) and the first South/North Human Genome Meeting in Brazil. Discussions in the latter meeting emphasised the value of genome diversity studies for involving developing countries in human genome research through an interest in their own populations and diseases.

The last of the three major planning workshops for the HGD Project took place in September 1993 in Sardinia. The meeting was organised by HUGO Europe and comprised three separate workshops. On 9th and 10th September there were two parallel workshops, one concerned specifically with the development of the European component of the global HGD Project and one concerned with issues of importance to the anthropologists involved in the project (extending the discussions begun in the second workshop at Pennsylvania State University). The main workshop, devoted to the global HGD Project, was held on 11th and 12th September with the workshops held on the previous two days providing input either in the form of written reports or through oral presentations.

The main workshop in Sardinia was attended by 80 people with very wide international participation. Twenty countries were represented including (in addition to the United States and European countries) China, India, Japan, South Africa, Kenya, Australia, Pakistan and Latin America. There was also broad participation from many disciplines, ranging from molecular genetics to anthropology. The meeting was funded from a variety of sources: The Porto Conte Research and Training Laboratories Foundation, Sardinia; the European Commission; the Soros Foundation; the United States National Science Foundation, National Institutes of Health and Department of Energy and HUGO Europe.

The consensus views of the workshop participants are summarised in the next three sections of this document under the broad headings: Scientific aspects of the HGD Project, Ethical aspects of the HGD Project, Management/Organisation of the HGD Project.

Participants in the international planning workshop for the HGD Project

9 - 12th September 1993, Sardinia, Italy

Dr Marcella Attimonelli (*Italy*)
Dr Ramiro Barrantes (*Costa Rica*)
Dr Jaume Bertranpetit (*Spain*)
Dr Julia Bodmer (*UK*)
Dr Bryan Bolton (*UK*)
Dr Dan Bradley (*Eire*)
Dr Anne Cambon-Thomsen (*France*)
Dr L Luca Cavalli-Sforza (*USA*)
Dr Michael Crawford (*USA*)
Dr Alan Doyle (*UK*)
Dr Georgia Dunston (*USA*)
Dr Aldur Eriksson (*Netherlands*)
Dr Marc Feldman (*USA*)
Dr Anne-Mare Frischauf (*UK*)
Dr Eugene Ginter (*Russia*)
Dr Henry Greely (*USA*)
Dr Takafumi Ishida (*Japan*)
Dr Juhe Kere (*Finland*)
Dr Judith Kidd (*USA*)
Dr Nikolay Kolchanov (*Russia*)
Dr André Langaney (*Switzerland*)
Dr Partha Majumder (*India*)
Dr Qasim Mehdi (*Pakistan*)
Dr Luciano Milanesi (*Italy*)
Dr Onesmo ole-MoiYoi (*Kenya*)
Dr Richard Mulivor (*USA*)
Dr Barbara Parodi (*Italy*)
Dr Vania Prado (*Brazil*)
Dr Colin Renfrew (*UK*)
Dr Massimo Romani (*Italy*)
Dr Alicia Sanchez-Mazas (*Switzerland*)
Dr Leo Schalkwyk (*UK*)
Dr Eric Seboune (*France*)
Dr Marcello Siniscalco (*Italy*)
Dr Chris Stringer (*UK*)
Dr M I Voevoda (*Russia*)
Dr Ryk Ward (*USA*)
Dr Mark Weiss (*USA*)
Dr Edward Whitehead (*EC*)
Dr Francesco Baralle (*Italy*)
Dr A S Santachiara Benerecetti (*Italy*)
Dr Anne Bowcock (*USA*)
Dr Walter Bodmer (*UK*)
Dr Batsheva Bonné-Tamir (*Israel*)
Dr Stephen Bryant (*UK*)
Dr Daniela Calò (*Italy*)
Dr Mauro Congia (*Italy*)
Dr Francesca Cucca (*Italy*)
Dr Gaby Dover (*UK*)
Dr Henry Erlich (*USA*)
Dr Liz Evans (HUGO)
Dr Jonathon Friedlaender (NSF, *USA*)
Dr Charles Gardner (Fogarty Institute, *USA*)
Dr Takashi Gojobori (*Japan*)
Dr Bernd Herrmann (*Germany*)
Dr Trefor Jenkins (*South Africa*)
Dr Kenneth Kidd (*USA*)
Dr Andreas Klepsch (EC)
Dr Doron Lancet (*Israel*)
Dr Lucio Luzzatto (*UK*)
Dr Thomas G Marr (*USA*)
Dr Tommy Meo (*France*)
Dr Guido Modiano (*Italy*)
Dr John Moore (*USA*)
Dr Svante Pääbo (*Germany*)
Dr Alberto Piazza (*Italy*)
Dr Terry Ray (*UK*)
Dr Otto Ritter (*Germany*)
Dr Aida Romashenko (*Russia*)
Dr Ralph Santos (*Italy*)
Dr Rosaria Scozzari (*Italy*)
Dr Sue Sarjeantson (*Australia*)
Dr Montgomery Slatkin (*USA*)
Dr Bryan Sykes (*UK*)
Dr S Wang (*USA*)
Dr Ken Weiss (*USA*)
Dr Stephen Whitehead (*Eire*)
Dr Gian-Luigi Zanetti (*Italy*)

V. SCIENTIFIC ASPECTS OF THE HGD PROJECT

Collection of Samples

i. Choosing populations to be sampled

To create a resource that can be considered to reflect the genetic variation in populations worldwide, choices must be made. Using language as a criterion, the world contains over 5,000 populations with distinct properties and possibly distinct gene frequencies. Many of these populations are very large and geographically dispersed and multiple sampling will be needed to provide adequate representation. It is simply not feasible to study all populations in detail within the HGD Project and representative populations therefore need to be selected.

In considering populations that might be included in the study, participants in the planning workshops have been well aware that, whatever method is used to select populations, the result will be imperfect, almost certainly controversial and, at worst, divisive. For that reason, no definitive lists of populations to be included - or not included - in the HGD Project have been produced. The discussions which took place in the second planning workshop held at Pennsylvania State University in October 1992 (specifically concerned with selection of representative populations) and which were extended in the workshop in Sardinia, have concentrated instead on the criteria for defining the categories of populations that should be sampled within each region of the world. Lists of examples of populations that fall within these categories have been drawn up in the workshops but these are only examples. Other examples would be equally appropriate.

ii. Categories of Populations

The list given below defines categories of populations that should be sampled for the HGD Project in terms of some, but by no means all, of the questions that can potentially be answered. It is not exhaustive and it will undoubtedly be modified as the Project develops in order to meet practical constraints, changes of priority and increasing knowledge and suggestions by the populations themselves. The choice of specific populations therefore remains an open issue, but all major regions of the world should be represented and each region should define its own local sampling issues.

Categories of populations:

1. **Populations that can answer specific questions concerning the processes that have had a major impact on the genetic composition of**

contemporary 'ethnic groups', language groups and cultures. Such questions include, but are not restricted to, the following: the origin of New World populations; the Bantu expansion; genetic consequences of the Indo-European expansion in Europe and in the Indian sub-continent; the genetic consequences of population expansions and migrations that followed the development of critical cultural advances, such as the domestication of plants and animals. Ancient human material is also likely to prove relevant to such questions.

2. **Populations that are anthropologically unique,** exhibiting unique cultural or linguistic attributes that distinguish them from their immediate neighbours. Sampling such populations may help resolve local anthropological and archaeological questions.

3. **Populations that constitute linguistic isolates.** The genetic study of such populations may shed light on the degree to which cultural evolution parallels the processes leading to genetic microdifferentiation.

4. **Populations that might be especially informative in identifying the genetic etiology of important disease.** One example would be studying Siberian populations to determine whether they manifest any attributes of the susceptibilities of Native Americans to diabetes.

5. **Populations that are in danger of losing their identity as genetic units.** Such populations have the potential to reveal unique constellations of genetic variability, before this information is lost forever. By the same token, there is some justification for studying populations whose cultural or linguistic integrity is threatened, although such populations may not be given the same priority as populations whose genetic identity is endangered.

iii. Criteria for selecting populations

In addition to the considerations listed above, **much of the information generated within the HGD Project will only be useful if it throws light on the questions about human history and identity that concern geneticists, anthropologists and historians or the questions that are raised by populations themselves.** These questions are of various different types:

Statistical questions

Questions that might be studied include:

1 What are the frequencies of various genes in different populations?

2 What are the levels of genetic heterozygosity or diversity within and among populations?

3 What special phenotypes, such as diseases, characterise certain populations, and are these correlated with genes or genotypes?

4 How do within- and among-population statistics compare? What is the variation among various human populations?

5 How could genetic and linguistic differentiation be compared and statistically tested?

Genome-wide questions

Examples of questions that fall into this category are:

1. What is the history of human population size? It is obvious that the human population has grown greatly but the pattern is not known. Were there bottlenecks or sudden expansions? Pairwise comparisons of neutral alleles in mitochondrial DNA in a worldwide sample suggest a rapid expansion in the Pleistocene period. This and similar hypotheses can be tested more completely with nuclear genes.

2 What is the history of migration and population subdivision?

3 What is the relative importance of random drift and selection? It has been realised that both have been important but their relative roles still needs to be assessed. Neutral alleles are particularly useful for phylogenetic inferences. There are a number of specific questions such as the importance of selection for structural or regulatory genes.

4 Is there spatial patterning among populations? Are populations that are geographically close together more closely related? What are the correlations of genotype and distance? Are there geographical clines and, if so, why?

Locus-specific questions

Two key questions are:

1 How is disease susceptibility distributed in and among populations? What accounts for important genotype-phenotype relationships and for genotype-by-environment interactions within and among populations?

Case-control studies of alleles and disease would require more sampling and differently targeted sampling than is anticipated for the HGD Project. Nevertheless, some geographical-epidemiological surveys would be feasible. For example, it would be possible to

evaluate variation at a locus of known function across populations or correlations of some easily observed phenotypes with candidate loci among populations.

2 What are the relationships among genetic, cultural, linguistic and ecological variables?

When many genetic markers are available, some analyses can be done with only 25 individual samples from each population, although a larger sample size is obviously more desirable. Examples are the analysis of principal components of allele frequencies and interpretation of these genetic components with respect to cultural, linguistic and geographical variables.

Other questions

Other questions of interest are not likely to be answered solely from the data generated within the HGD Project. However, combining these data with that from other studies will be very useful in considering questions such as:

1 How do mutation and recombination rates vary among populations and among loci?

2 How much linkage disequilibrium is there in the population and how is it influenced by stratification and epistasis?

iv. General approaches to collecting samples

In previous years, samples from a large number of different populations have been collected by individual researchers. As a result, cell lines as well as DNA samples exist in many laboratories around the world. Consideration should be given to assembling the most relevant of these samples into the central repositories for the HGD project in order to give researchers access to them.

A number of different approaches could be used to build up the central repositories of biological samples for the HGD Project. These include:

a) *Sampling de novo specifically for the HGD Project.* Populations would be selected specifically for the HGD Project and a sampling expedition set up.

b) *Opportunistic sampling.* Where populations appropriate to the HGD Project are already being investigated for some 'non-genetic' purpose, the provision of small additional funding may enable sampling for the HGD Project to be undertaken at the same time.

c) *Using existing samples.* Some samples which already exist or are currently being collected, would be suitable for being deposited in the central HGD repository. One example would be the cell lines

being established as part of the 12th HLA Anthropology Workshop, to be held in Paris in 1996. The samples from this Workshop are likely to have special relevance for the HGD Project since they will include samples from such populations as African-Americans and American Hispanics.

Some substantial collections of serum, red cell lysates and other biological materials exist from which DNA could be prepared and used for HGD studies. However, better methods of extracting DNA from such materials would first be needed to ensure that not even small amounts of non-renewable DNA are irrevocably lost. A number of laboratories also have stocks of DNA that could be made available to the HGD Project under appropriate conditions.

v. Sampling strategies

Overall, the specific sampling strategy that is used to collect biological material from any population will depend on the primary scientific question that has motivated the inclusion of that population in the project. However, every effort will be made to ensure that the samples fulfill minimum criteria with respect to size and adequacy in order to ensure that the material deposited in the central HGD repositories is representative of the genetic diversity present in the population being sampled.

In general, as many individuals as possible should be sampled in each population. However, for some phylogenetic purposes, 25 individual samples may be sufficient, provided that an adequate number of genetic markers are evaluated for each (eg. markers from 100 - 200 different positions on the DNA). For many other purposes, such as investigating genetic microdifferentiation, evaluating haplotypes and other aspects of genomic evolution, it would be more appropriate to sample 100 - 200 individuals and **a norm of 150 samples is generally recommended.** For studies that investigate phenotype-genotype associations, sample sizes may have to be even larger.

Within these constraints, a large number of different sampling strategies could be contemplated, depending on the geographical region being examined. These include:

a) The use of a grid approach for populations that are relatively large and geographically dispersed. Such populations might include the large tribal populations of India (eg. the Bhils and the Gonds) and perhaps the Cree populations of Sub-Arctic North America.

b) Intensive sampling of single communities in order to study microdifferentiation within the population.

c) Deliberate sampling of groups of relatives to answer questions that relate to the segregation of disease susceptibility genes, the reconstruction of complex haplotypes, the transmission of specific alleles, or genomic elements.

d) At the population level, contiguous groups could be sampled in a fairly intensive fashion in order to answer questions that relate to the maintenance of barriers for gene flow (eg. genetic isolates and their neighbours). Conversely, the tempo and mode of gene flow between populations that differ linguistically could be examined by sampling a series of contiguous groups.

e) A stratified, proportional sampling approach based on ethnohistorical data, could be used for large contemporary populations. Such a strategy might be especially useful for populations whose genetic constitution derives from a mixture of formerly disparate groups (eg. African-Americans).

vi. Collection of socio-demographic data

In addition to the biological samples (of blood, hair or cheek scrapings), it is important that a minimum set of socio-demographic data is collected in a consistent way from all individuals and populations. To ensure that all individuals contributing samples to the HGD project's central repositories and database remain anonymous, each sample will be given a unique identification rather than being recorded by the name of the individual. Information about the name of the individual and the names of the individual's parents will be routinely collected in the field but this information will be treated as confidential and not released to the central database. Discussions are continuing about the extent of detail that will need to be recorded but the type of information is as follows:

the unique identification code
date of sampling
quantity of blood
existence of cell lines
extraction of DNA
deposition of DNA, etc.

In addition to this information, the following data should be recorded:

sex
age (or approximate year of birth)
current residence
place of birth
linguistic affiliation.

The following information about the individual's biological parents should also be collected wherever possible:

current residence
place of birth
cultural affiliation
linguistic affiliation.

Since there are numerous ways of defining and recording cultural affiliation, the exact manner in which this is done will vary from one group to another.

It may also be useful to collect social and cultural data about the individual but the exact way of coding this information will vary from one group to another and will have to be agreed by the group itself.

In addition to the above items, the investigator collecting the sample should also attempt to obtain specific information (latitude and longitude) about the geographic location of the community. However, this information may have to be treated as confidential.

vii. Field work issues

1 **As far as possible, the initial collection and handling of samples for the HGD Project should be done, in all regions of the world, by local investigators.** It is hoped that networks of reliable researchers can be established in each region to take responsibility for this and to assist in the general organisation and development of HGD activities in that region. It is essential that, in all regions, the collection and handling procedures conform to the guidelines and requirements of the HGD project. Where local facilities or expertise is limited, collaborations will be arranged to ensure that appropriate resources are available. As discussed later in this document, an essential feature of the HGD project is its potential for the transfer of 'enabling technology' to countries which have not yet become actively involved in human genome research.

2 **Since blood sampling is a mildly invasive procedure that has certain risks, appropriately trained people must be involved in collecting such specimens.**

3 **Those collecting samples for the HGD Project must respect cultural beliefs that may influence the type of sample to be collected** (eg. the sanctity of hair, or of blood, may be sufficient to preclude sampling that particular biological material from certain communities).

4 **It is important that there is willing and informed participation in the HGD project and the consent of individuals to participate must be obtained in a culturally appropriate manner.** People need to be informed about the overall objectives of the project in a manner they can understand within their cultural context. Ideally, participation in the sampling process will result in the individual subjects and the entire community becoming partners in the HGD endeavour.

5 **The customs and traditions of participating communities must be respected at all times. This applies particularly to extremely remote populations whose customs and traditions may be least familiar to the investigator.** The interest and support of the community must be obtained before sampling begins and the help of local contacts who are known and trusted by the community should be sought in obtaining information about local customs and beliefs. The kinship

system of the community must be fully understood so that biological and social relationships can be disentangled.

6 **Before samples are collected for the HGD Project, any necessary approval (for example, from governmental agencies) must be obtained.** For many of the more remote populations it is likely that approval will be required from a series of agencies. In gaining such approval there should be sensitivity to any possible conflicts of interest between the governmental agency and the community being sampled. The highest priority should be given to respecting the rights and wishes of the community being sampled - even if this results in abandoning plans to sample a particular community.

7 The existence of HIV/AIDS is a salutary reminder that, **when collecting biological specimens, there is a responsibility to both donors and investigators that appropriate measures be taken to minimise the risk of infection by pathogens.** This includes taking prophylactic measures against risk of infection by hepatitis and other pathogens. It is important that all involved in sampling is understand that DNA (to be extracted from samples) is non-infectious and that EBV cell lines (to be established from some samples) almost certainly do not carry HIV.

Long-Term Storage of Samples

i. Regional collecting centres

As indicated in the preceding section, **it is intended that most of the initial collecting and handling of samples for the HGD Project will be done by local investigators in each region of the world. It follows that one or more collecting centres must be established in each region to receive and store materials from populations in that region and to establish cell lines. Samples and cell lines (in amounts yet to be decided) would then be transferred from the regional collecting centres to the central repositories for the HGD Project to create the global resource.**

Whether in regional centre or central repositories, serious consideration must be given to the criteria for storage since it is intended to preserve samples indefinitely. Different types of samples will require different storage conditions. For example, cell lines should be stored in liquid nitrogen freezers but purified DNA samples, phage libraries, PCR amplified material, hair roots, cheek swabs, dried blood spots and isolated (viable or non-viable) white cells will all require different storage conditions for long term preservation.

With the process having been perfected by several billion years of natural selection, *in vivo* replication of DNA is the only mechanism known to amplify all aspects of an individual's DNA equally and faithfully. Many years of experience have shown that B lymphocytes transformed with the Epstein-Barr virus (EBV) are easy to culture and that the integrity of their DNA can be maintained through very large numbers of cell divisions. Of course, mutations do occur so it is important to try to maintain some cells from an early, still polyclonal phase of the culture, but experience with cell lines such as those maintained at CEPH (Centre d'Etude du Polymorphisme Humain, Paris) shows that mutation is not a major problem even over long periods of time.

The construction of DNA libraries as a means of preserving and amplifying the DNA extracted from collected samples may become an important aspect of the HGD project. Such libraries could be constructed in phage (or other vectors) from freshly-extracted DNA, from size-selected DNA fragments ligated to oligonucleotides used for PCR amplification or from the hybrid of DNA initially amplified by PCR (eg. with a mixture of random primer sequences, and subsequently cloned into phage for additional amplification). It is now well recognised that in all such libraries some of the genome is not preserved and that the organisation of the DNA in large segments is also lost. There will also be drift over successive replications of the material: differential replication results in significant loss over time of a fraction of the genome preserved in the library. Despite these limitations, the relative ease with which the majority of the genome can be preserved and amplified many fold makes the use of libraries attractive. As discussed later, techniques need to be developed for maximising the fraction of the genome that can be preserved and amplified indefinitely and for increasing the ease with which this can be done for a large number of individuals.

ii. Central repositories

Although much of the initial work of the HGD Project will be undertaken in different regions of the world, using local facilities, it is desirable that samples and data be stored in central repositories. It is recognised that the feasibility of supplying material and information to central repositories will vary in different regions of the world but the aim of the project developers is that samples and data collected under the auspices of the Human Genome Diversity Project be open to the scientific community. Open access may be used as a criterion in determining whether a sampling proposal should be recognised as part of the HGD Project.

A minimum of two central repositories in different locations is considered necessary so that the global resource is housed in duplicate and thus safeguarded against wholescale disaster in either one of the locations.

Many of the specific issues concerning the creation of central repositories about which decisions will have to made in due course have already been identified. These include the number and type of samples to be stored, the number of cell lines to be established and sustained, the number to be transformed with EBV, and the likely costs associated with transformation and with long-term storage of different forms of samples. However, further details are needed before decisions can be made about any of these. In particular, without an extensive description of the services required from the central repositories, any pricing can only be tentative.

iii. Access to the resource

Material and information kept centrally should be accessible to the scientific community. Details of access to central cell repositories and databases have yet to be worked out but it is likely that priority in availability of material would initially be given to those who agree to characterise the DNA for the bank of core markers against which all samples will be typed. Some laboratories may, for a fee, choose to offer a typing service to other laboratories not inclined or able to do this testing themselves. Researchers interested in questions unrelated to the HGD Project may be required to justify their access to the resource and provide reimbursement for expenses. All who use the resource to generate genetic information, whether or not on the core markers, will be required to file their results for inclusion in the HGD database.

Analysis of Samples

i. Types and testing of markers

In order to be able to define the relationships between populations, it is proposed that all samples collected and studied within the HGD Project are tested against an agreed core set of alleles or other genetic markers. In addition to these core markers, a wide variety of other markers will eventually be studied, including those that have clinical relevance or a particular relevance to regional subsets of populations.

The choice of core markers is one of the most important stages in the initial phase of the project and, in the planning workshops, consideration has been given to the types of markers that are currently available and their likely usefulness for the project. In addition to the technical limitations on typing methods the issues involved include the need to distinguish between markers from functional and non-functional regions of the genome. Many of the available markers, such as repeat markers (micro- and mini-satellites) and most RFLPs (see below), are found in introns or in the regions flanking genes, areas which are either unlikely to be functional or known to be non-functional. The variation in these markers, which should be neutral selectively, will need to be compared with variation in known functional genes. It is the latter, mainly amino acid substitutions (ie. mis-sense changes), which are likely to have some selective effects or be associated with interesting phenotypes such as facial differences or skin colour.

The usefulness for the HGD Project of markers currently available can be summarised as follows:

Classical markers

A lot of useful information about populations has been obtained using classical markers (such as the ABO blood grouping). To ensure that such data can be integrated with future studies of human genome diversity, these markers need to be converted to PCR-based systems (see below). Where this has not yet happened, PCR-based systems should be developed.

RFLP markers

Hundreds of RFLP (restriction fragment length polymorphism) systems, each characterised by a probe/enzyme combination, have been developed and already applied extensively to population studies. However, although RFLPs were initially the most useful markers, there are technical limitations on their use in the HGD Project. RFLP analysis requires relatively large amounts of high molecular weight DNA per test and therefore in general requires the availability of lymphoblastoid cell lines. In addition, the analysis is not readily amenable to a multiplex approach. It is PCR-based typing systems (into which many of the RFLPs could be converted) will be much more convenient for large scale studies.

Minisatellite markers

Minisatellite markers are loci consisting of tandem repeats of small basepair units giving alleles which range from 0.5 to 30kb in length. They are also known as VNTR (variable number of tandem repeats) loci Hundreds of such loci have been mapped but, for several reasons (such as the difficulty of classifying alleles precisely due to microheterogeneity in repeats), they are considered to have limited applicability for population studies.

Y chromosomal polymorphisms

Y chromosomal polymorphisms would be of great interest in contributing to male lineages and male patterns of migration in a complementary manner to that for mitochondrial DNA which corresponds to female lineages.

Microsatellite loci

These loci, consisting typically of 10-30 repeats of a 1-5bp repeat sequence, are abundant in the genome. Thousands of them have been isolated, in particular $(CA)_n$ loci, and many have been mapped. They are multi-allelic (1-20 alleles per locus, usually about 5), highly informative and generally have low mutation rates. Potentially, microsatellite loci hold major advantages for studies of human genome diversity, including the following:

- The allele length (repeat copy number) variability can be detected.
- Ancient DNA (from bones, for example) has already been successfully typed using such loci.
- Forensic DNA typing laboratories are actively developing such loci for forensic use and have already invested substantial effort in identifying appropriate loci and developing standardised typing formats and reference allele ladders; such standardised markers would provide the HGD Project with an excellent resource.

It is therefore strongly recommended that microsatellite loci, and in particular the more easily typeable tri- and tetranucleotide repeat loci, should be incorporated within standard sets of markers.

UsingPCR-based systems to type polymorphic markers

For typing polymorphic markers, PCR-based systems have a number of major advantages. Most typing can be done on 1-10ng of genome DNA so 1ml of blood will yield sufficient DNA for 1,000 - 10,000 marker tests. Typing is probe-independent and defined only by oligonucleotide primers; it is therefore portable. The typing is fast, can be conducted in multiplex format and is amenable to automation. PCR-based systems are also highly flexible, allowing different

methods for accessing the same polymorphism. **It is therefore recommended that all DNA polymorphisms studied in the HGD Project should be typed by PCR-based methods.**

Digital DNA typing

Digital DNA typing assesses variation in sequence between repeat units at minisatellite loci and can be applied not only to genomic DNA but also, in rinciple, to whole genome PCR libraries. This method allows minisatellite alleles to be defined with precision on the basis of internal allele structure and, at the more variable loci, can reveal huge levels of allele diversity (10^8 alleles or more world-wide). Groups of closely related alleles sharing regions of structural similarity can be identified and appear to show substantial population specificity. The possible uses of this novel approach for the HGD Project should be explored.

ii. Choice of markers

In addition to the considerations reported in the previous section, it is important to define other properties of the core markers which will be adopted by the HGD Project and for which all samples would be typed. These properties fall into a number of categories:

Technical

Clearly any markers must be technically robust in the sense that results are reproducible and unambiguous. The process for evaluating potential core marker systems should include a quality control element to ensure this robustness both within and between laboratories.

As far as possible, marker systems which could be used in regional and local laboratories should be adopted to enable these laboratories to participate fully in the analytical side of the project. To ensure this, it may be necessary to develop low-tech versions of some marker systems for use in the regions at the same time as effort is being put into sophisticated automated typing in high-tech reference centres.

Where possible, detection systems should be base on non-radio-active components such as silver-staining or colour reactive ligands.

DNA-based systems

It is essential that all core markers are available in DNA-based versions (almost certainly incorporating PCR amplification - see previous section), as the material distributed from the central repositories will be in the form of small amounts of DNA. Regional collecting centres may use more traditional methods for classical markers in the core set on locally collected material.

Informativeness

Clearly the usefulness of any marker depends on its ability to maximise the information obtainable from the samples in the HGD collection - which is different from saying that only those marker loci having the highest number of alleles should be considered. Some loci may be too variable for useful comparisons of distantly related populations while being well-suited for more local differentiation.

Lineage markers

Both single (mitochondrial and Y-chromosome) and mixed (nuclear) lineage markers will be useful for the project. The anthropological interest in comparisons between different lineage markers will be considerable if they could help decide the sex ratio of past migrations.

Evaluation

Before being adopted for the project, any newly developed marker system will need to be evaluated for the different criteria considered above. A small reference collection will be established for this purpose.

From all their discussions that have taken place about markers, it is concluded that a small set of markers is sufficiently well established to be adopted as core markers without further evaluation. These markers include certain HLA alleles, mtDNA control region sequencing and some blood groups (ABO, MN). A second group will be adopted when available as PCR versions and these might include RFLPs where high between-population variabilities (F_{st} values) have already been established using other techniques. Markers used by the forensic community have undergone rigorous technical evaluation and could be added to the list if their usefulness in this project is demonstrated by the pilot study on the refence collection. Further markers will be added to the list after successful evaluation.

iii. Development of technology

In order to carry out the large scale survey of human genome diversity proposed by the HGD Project, technical developments will be needed in a number of areas including the following:

1. **Production of cell lines from the collected samples**
2. **Production of an indefinite supply of DNA from samples using PCR-based techniques**

The discussions of the developments needed in these areas are summarised below:

Production of cell lines from the collected samples

Much of the work to be carried out within the HGD Project will be done with extracted and PCR-amplified DNA. However, the standard approach

of restriction endonuclease digestion of blood DNA followed by ligation of oligonucleotide linkers and whole genome PCR creates amplifiable 'libraries' which are non-representative, large DNA fragments becoming progressively lost during repeated amplification. It is therefore proposed that cell lines are developed from some of the collected samples in order to sustain a supply of 'original sequence' DNA.

The usual way of establishing cell lines is by transforming lymphocytes with EB virus and a high efficiency of transformation can be achieved with the current methodology provided that blood samples can reach the laboratory within a few days of collection. For some of the samples to be collected within the HGD Project this may not be possible so improved techniques for transforming and preserving lymphocytes are needed as well as new approaches to the development of cell lines. In order to minimise the necessity for collecting blood, these new approaches should include the development of cell lines from cells other than lymphocytes - for example, the epithelial cells obtained from a cheek scraping or the epithelial cells and fibroblasts contained in the roots of a hair. Techniques need to be worked out for sterile, short-term propagation of cells from such sources and for the possibility of efficient transformations using one or more of the oncogene and oncogenic virus-based systems that have recently been developed in other contexts.

Indefinite production of DNA using PCR based techniques

The production, maintenance and storage of cell lines is expensive and it is therefore desirable that alternative ways of providing an indefinitely-renewable source of DNA from collected samples are also developed. The only alternative approach that is feasible at present is that of amplifying the whole genome using PCR technology. However, as previously mentioned, DNA sequence is not faithfully reproduced over many copies using current PCR technology. Further research is therefore needed, particularly in the following areas:

- Development of new methods for making representative PCR 'libraries' such as by the sonication of genomic DNA followed by size selection of an appropriate narrow size range of DNA fragments, addition of linkers and then whole genome PCR.

- Development of criteria for validation of PCR libraries. Standard protocols need to be developed for library re-amplification; representativeness should be tested using a battery of polymorphic marker loci and stability of representation by re-amplification; the largest target DNA that can be faithfully tested on such a library should be characterised.

- Generation of duplicated banks of representative DNA libraries (for example, in microtitre plate format) that can be distributed to laboratories participating in the HGD Project.
- Generation of a database of DNA 'fingerprints' from each genomic DNA and its corresponding PCR library (using

microstaellite loci or digital DNA typing, for example). These could be made available to all participating laboratories to ensure the correct identity of all samples tested.

- Using PCR it is possible to amplify specific DNA segments from extremely small sources such as single cells and the minute samples available in forensic and archaeological work. This has opened up new perspectives in several fields and, suitably developed, could be of great value to the HGD Project. The ability to replicate DNA from very limited amounts of material could lead to the development of highly representative libraries from blood or serum samples that have been stored for many years, thus reducing the need for new samples to be collected. At present, a major limitation of this approach is that the number of amplifications that can be performed from one sample is limited to one or very few. To overcome this, technological advances are needed in the ability to propagate random nucleic acid fragments in vitro from such samples

iv. Training and technology transfer

From the outset the HGD Project has been planned on a global basis with world-wide coordination of sampling and testing. Local participation in all areas of the world will be essential and the success of the project will also be entirely dependent on international collaboration and cooperation.

It is recognised that not all regions of the world are experienced in the techniques of molecular biology and genetics and that some countries will not, in the foreseeable future, acquire the 'cutting edge' technology that is needed for the mapping and sequencing work of the Human Genome Project. However, it is feasible for the more limited technological demands of the HGD Project to be met by most countries, given training of laboratory staff and help with techniques. **One of the exciting aspects of the HGD Project is that it offers all countries a unique opportunity to become involved in, and contribute to, the global human genome initiative by undertaking the collecting and typing of samples from their own region as well as other studies of local interest.**

As mentioned previously in this document, the intention is that as much of the initial work of the HGD Project as possible, in all regions of the world, is carried out by local investigators. **One important aim within the project will be to help establish or upgrade laboratories in many parts of the world to enable greater local participation.** There are numerous ways of aiming to achieve the transfer of appropriate tranining and technology. One is through specific collaborations between scientists in different countries, which could be encouraged and facilitated by the regional HGD committees in accordance with identified local needs.

Another is to take advantage of the networks of laboratories that have already been set up in different regions of the world under the auspices of, for example, the International Centre for Genetic Engineering and Biotechnology (the ICGEB, which has main centres in Trieste, Italy and New Delhi, India) or the international HLA workshops. All such avenues will be explored in detail, in collaboration with funding agencies, as the HGD Project progresses. The intention is not merely to train and support local laboratories solely for a through-put processing role in the project - the project cannot be a worldwide success and benefit all peoples unless permanent resources are established in each of the different regions. Development of the regional collecting centres will therefore be very important to enable them to play an ongoing and increasing role in the analysis of data from each region.

v. Pilot Project

At the HGD planning workshop held in Sardinia in September 1993 it was agreed that, **while further details of the HGD Project were still being developed, a small pilot study should be undertaken to begin to test the suitability of markers.** This pilot study will focus on testing many existing and new DNA polymorphisms using a panel of about 200 cell lines from about 20 populations well-distributed around the world. Four main objectives for the pilot study have been identified:

- to assemble an initial small set of cell lines very broadly representative of global genome diversity to serve as a pilot for a variety of organisational procedures that will be important to the full HGD Project.

- for this set of cell lines to provide a panel of DNAs on which to test a large number of diverse genetic markers in order to evaluate their suitability for inclusion among the initial core set of markers used in the HGD Project.

- for this set of cell lines to be a panel of DNAs on which future markers can be tested and evaluated.

- to provide a dataset of marker typings from multiple laboratories on a single set of DNA samples on which database systems and data management can be piloted.

It was proposed that the pilot study should not involve the collection of any new samples but that existing cell lines should be used as the source of DNA. The amount of DNA required is not expected to be large since the pilot project will, as far as possible, use PCR-based typing.

As far as is feasible, The laboratories participating in the pilot project will exchange markers so that duplicate typing is carried out. This will allow the reliability of techniques to be evaluated as well as determining the ease with which various markers can be typed. As the pilot project - like the entire HGD Project - is a collaborative effort, all contributors of cell lines and all those participating in the marker typing will be required to ensure that the data they are responsible for collecting is complete and is that it is deposited in the central repository.

The Executive Committee of the HGD Project will decide on initial analyses and procedures for evaluating data. One obvious analysis will be to evaluate F_{st} for each polymorphism to be used in the core set since those markers with a higher global F_{st} will allow a finer scale of evaluation of population relationships. Other types of analysis are being discussed. A database suitable for the pilot study will be developed from an existing small laboratory-specific database. This database and the data from the pilot study will then serve as the testbed for the specification and design of the database systems required for the full project. The results from the project will be coordinated and made publically available in a manner that will serve as a pilot for data access for the full HGD Project.

Development of Database

The global resource that will be created by the HGD project will exist not only as a collection of biological samples that represents the genetic variation in human populations but also as an open, long-term, genetic and statistical database on variation in the human species that will accumulate as the biological samples are studied by scientists from around the world.

The proposed project is complex and far-reaching and will generate a very large amount of data. It will therefore be increasingly important that the data collected by many different laboratories is brought together into one or more central databases with public access so that any researcher involved in any study of variation in the human genome can obtain the necessary raw genetic data. **It remains to be decided whether a single database should include all the information relating to the project (eg. genetic, historical, linguistic) or whether, as seems more likely, a series of separate, specialised databases will be developed.** In the latter case, integrating the different databases with each other and ensuring full compatibility and transferability of data between them will be an imporant part of the project. It will also be important to ensure a high degree of transferability of data between the databases used in the HGD Project and a variety of other databases containing information about the human genome. Linking to the Genome DataBase (GDB), the main genome database for the Human Genome Project, will be particularly important.

It is recognised that there will be many major practical problems concerned with data submission, access and analysis. To ensure that these problems are given detailed consideration as early as possible in the development of the project, an expert 'Database subcommittee' of the HGD Executive Committee has been established.

The consensus view is that the HGD Project will require a single data coordinating centre with responsibility both as a data repository and as a site for software and data distribution. It is envisaged that the regional collecting centres, acting in concert with the appropriate regional committees for the HGD Project (see later), could act as local points of data submission. The regional collecting centres would then be responsible for passing data on to the data coordinating centre and also for facilitating local access to the data and relevant analytical software. The HGD data coordinating centre should not need computing and human resources on the scale required by the Genome DataBase (GDB) and could well be associated with existing initiatives but funding would be needed for the coordination and support of the regional centres.

One of the early goals of the informatics component of the HGD Project should be to identify areas of overlap with existing database resources and to consider whether their organisational models can profitably be transferred to the project. For example, it would obviously be valuable to draw on the experience with GDB. In addition, the model of regional data collection proposed for the HGD Project has similarities with the EC-funded EUROGEM programme that is concerned with increasing the density of markers in the CEPH database. It is even closer, as a model, to the way it is felt that the computing infrastructure for the next HLA workshop (to be held in Paris in 1996) should be organised. The experiences gained in setting up the anthropological component of the next HLA workshop will undoubtedly be of great importance in assessing the nature and extent of problems in data accessibility and submission in regions targeted by the HGD Project.

Another of the working goals of the HGD Project should be the transfer of information via the international Internet, which has already become the dominant form of communication

between Institutions and individuals involved in the Human Genome Project. As some regions of the world do not yet have a well-established infrastructure supporting network communications, this proposal might seem unworkable. However, there are recent, inexpensive technological developments which can be used to circumvent what might appear to be insurmountable difficulties. These include data transfer over short distances by radio and over longer distances by tapping into low-level satellites. Such developments are not yet widely used but the data coordinating centre for the HGD Project, in conjunction with the Database subcommittee for the project should be regarded by the technical professionals in the regional collecting centres as a source of expert advice and support.

There has not yet been detailed discussion of the primary data that should be collected and stored beyond that already mentioned under 'Field Issues'. Consideration of such issues will form part of the work of the Database subcommittee. However, the small-scale pilot project will be of major value in helping to specify the design and parameters of the database system required for the full HGD Project.

VI ETHICAL ISSUES

A considerable amount of time in the HGD planning workshops has been devoted to reviewing the ethical issues involved in the proposed project. The areas of concern range from the preservation of individual rights within indigenous communities, where the presumption of 'informed consent' and adherence to 'Western ethics' is likely to be at variance with common practice, to a concern with the preservation of intellectual property rights. The following guidelines are proposed:

Proposed Guidelines

1 **The HGD Project and its participating researchers must always respect the humanity of the sampled individuals and the cultural integrity of the sampled populations.** This respect demands that collections proceed only with the knowledgeable agreement of both the population and members. It also demands that the project takes a primary responsibility to avoid harming sampled individuals or their communities. Wherever possible, studies should be carried out by local investigators known to and trusted by the population to be sampled and from the country in which they live.

2 **Informed consent is both an ethical imperative and a legal requirement. The HGD Project must satisfy both conditions.** To do so, the question of obtaining informed consent from participating individuals cannot be considered a mere formality but must be obtained in a culturally appropriate manner. This may differ from country to country. In addition, when scientists from one country are funded to undertake the collecting of samples in another country, there may be significant differences regarding the obtaining of informed consent between the funding country and the country in which the samples are collected. Funding agencies should respect these differences and not seek to impose their own cultural procedures. The requirement in all cases is for people to be informed both of the collection procedure and of the overall goals of the Project in ways they can understand or are appropriate to their culture. All participation should be voluntary. The objective should be to have the individual participants and the entire community become partners in the scientific effort. The idea of informed consent should also include an appropriate form of feedback of the results of the study to the sampled population.

3 **Researchers should actively seek ways in which participation in the HGD Project can bring benefits to the sampled individuals and their communities.** Examples of such benefit include health screening, medical treatment or educational resources.

4 **One way to avoid bringing harm to the sampled individuals or their communities is by protecting the confidentiality of those sampled and, in some cases, of their entire community.**

5 Although very unlikely, it is neverless possible that the results of the HGD Project may lead to the production of commercially beneficial pharmaceuticals or other products. **Should a patent be granted on any specific product, the project must work to ensure that the sampled populations benefit from the financial return from sales.**

6 Human history - and the human present - is full of racism, Xenophobia, hypernationalism, and other tragedies stemming from beliefs about human populations. In the past, some of those tragedies have been perpetrated by, or aided by, the misuse of scientific information. **All those involved in the HGD Project must accept a responsibility to strive, in every way possible, to avoid misuse of the project data.**

7 Many people in the world have, at best, a limited understanding of human genetics. Some fear the consequences of human genetic research, in part because of the limits of their understanding. To scientists involved in the HGD Project, such fears may not seem justified or even, in some cases, fully rational but **the concerns are very real to the people involved and they must be addressed. It is essential that a world-wide 'public awareness' programme is included within the project to educate people about its aims, methods and results.**

8 Inevitably, the ethical issues faced by the Project will evolve over time. **The issues must therefore be kept under continual review.** The widest possible consideration of the issues should be encouraged.

9 The transfer of technology to developing regions of the world, which is an integral part of the proposed project, should contribute positively to the development of self-sufficiency in these regions. **The help given should not be superficial and of only short-term usefulness.**

10 **There should be a feed-back of information to populations that participate in the HGD Project,** most especially about any aspect of the Project in which a particular interest had been expressed.

Major areas of concern

The HGD Executive Committee have identified four major areas of ethical concern. The Committee believes that these areas must be kept under constant review and the subject of continuing consideration.

a) Collection issues:

Respect for individuals and their cultural integrity must be the foundation on which all collection efforts are based. There are three general ethical concerns that relate to the actual collection procedure:

Informed Consent
The informed consent of all those participating in the HGD Project is vital. Regardless of the varying legal requirements that may need to be met, truly informed consent requires that people agreeing to participate understand i) that the actual collection of the sample involves some (specified) risks although these are very small, ii) that the sample collection will cause a little discomfort, and iii) that DNA from the sample that is given will be stored in a repository and may be used by many investigators for a long time period (for many subjects this also requires that they understand that cell lines will be established). Participants should also be given some understanding of the specific analytic goals of the investigator engaged in the sample collection, as well as the more general goals of the HGD Project. The overall analytic goals of the HGD should be described in general terms, as these are likely to change over time. This will ensure that samples can continue to be used as new research questions and techniques are developed.

Testing for disease
The issue of testing for disease is not only very important but also many faceted. For example, there are obligations to resolve in regard to testing for infectious disease, which raises issues of protecting lab workers and investigators, as well as issues of protecting the individuals from whom samples are collected. There are also obligations to resolve with regard to testing for non-infectious disease. In all cases, there are many questions to be addressed. For example, Is it ethical to test for *any* disease without providing pre-testing counselling or evaluation of the test? Who is to be informed of any results? If disease is tested for, what are the obligations for providing treatment?

Confidentiality
The anonymity of all individuals participating in the HGD Project must be preserved in order to provide protection against any possible abuse or adverse effects arising from the consequences of the study. For remote populations is may also be the case that the entire population wishes to have its anonymity preserved (or it may be politically necessary that the group not be identified). These wishes must be respected with the same diligence that individual anonymity is respected.

b) Intellectual Property Rights:

The patenting of any products which may be derived from the samples contributed to the HGD Project should include provision for the financial return on sales to benefit the sampled population or individual. There are too many precedents where this principle has been abused. (eg. seed banks and transgenic plants have been developed with

the assistance of indigenous populations and then sold back to them at high prices). In many areas of the world, such abuses have sensitized people to this problem. It must be made very clear that the HGD Project has no financial or commercial interest in the collection and analysis of the samples. In addition, it would be desirable to put the management of the database into the hands of an international (respected and independent) organisation. The HGD project is based on the fundamental principle that the resulting data can be accessed by any scientist and it is important that this principle is built into any sample collection.

c) Racism, Xenophobia and Hypernationalism

As world history includes a number of examples of racists either misusing genetic data or using the rhetoric of genetics without any real data, **every possible effort must be made to minimise any misinterpretation of the analysis and plans for the HGD Project.** Public attention should also be drawn to the fact that past studies on genetic diversity in human populations have actually shown that typological classification of humans into a small number of 'races' is scientifically invalid.

d) Public Understanding

In some parts of the world, fear and misunderstanding about genetics are facts of modern life and there is considerable potential for the HGD Project to be seriously undermined through the spreading of inaccurate or incomplete (and, therefore, biased) information. **Again, an active educational effort ongoing throughout the course of the Project is probably the most effective way to anticipate and address misunderstandings.** There are many different activities that could be undertaken within such an effort, including the holding of open fora to allow public comment. As a starting point, the HGD Executive Committee feel it would be helpful to establish a group of individuals who could respond immediately to misunderstandings and uninformed comments of the type that have already been published about the project in the lay press.

VII MANAGEMENT/ORGANISATION OF THE HGD PROJECT

Introduction

The international workshop held in Sardinia in September 1993 was the last of the initial series of planning workshops proposed by the HGD Committee established by HUGO on an *ad hoc* basis in 1991. The purpose of holding these workshops was to explore the major scientific issues involved in the proposed project and to define a framework for the project which would provide a basis for more detailed discussion and planning. The Committee had recognised from the outset that the successful implementation of the project would not only involve scientists but also various national and international non-scientific groups, all of which could and should contribute constructively to the project's development. However, it was also felt that, in view of the multi-faceted nature of the project, discussions involving *all* interested parties would only be fruitful if the scientific framework for the project was first defined. The Committee regrets that some non-scientific groups with a justifiable interest in the HGD Project have assumed that the planning of the project is much further advanced than is actually the case and have publicly and extensively criticised aspects of the project which have not yet even been planned let alone implemented.

At the workshop in Sardinia the proposals outlined below for the management/organisation of the HGD Project were presented to the participants, who represented all the regions of the world and included many of those who would be most actively involved with the project. After considerable discussion, the participants approved without dissent the proposed organisational structure. It was agreed that the proposal would be presented to the HUGO Council on their behalf by Sir Walter Bodmer. This was done at a meeting of the Council in November 1993 and formally approved by the Council at its meeting in January 1994.

Overall plan

1 **The HGD Project will be developed under the auspices of HUGO,** which initiated the detailed consideration of the proposed project. The involvement is thought appropriate as the success of the HGD Project will be entirely dependent on international collaboration and cooperation and HUGO already has a major coordinating role within the Human Genome Project. As part of its coordinating work, HUGO has a particular responsibility to ensure that the ethical issues arising from the global human genome initiative are addressed internationally, which will be an important consideration for the HGD Project.

2 **Any individual researcher, research group or laboratory who wishes to be recognised as contributing to the HGD Project must adhere to the**

guidelines outlined in this document. The HGD Executive Committee reserves the right to publicly dissociate from the Project any researcher, research group or laboratory which fails to meet this criterion.

2 The *ad hoc* HGD Committee, established under HUGO in 1991, will become the Executive Committee for the project and a standing committee of HUGO. The membership of the committee was extended at the workshop in Sardinia from ten to thirteen by the addition of three new members - Takashi Gojobori (Japan), Partha Majumder (India) and Onesmo ole-MoiYoi (Kenya) - in order to give more global representation. The membership of the committee is therefore as follows:

Dr Julia Bodmer (UK)
Dr Walter Bodmer (UK)
Dr Luca Cavalli-Sforza (USA) - Chairman
Dr Marc Feldman (USA)
Dr Takashi Godjobori (Japan)
Dr Ken Kidd (USA)
Dr Mary-Claire King (USA)
Dr Partha Majumder (India)
Dr Onesmo ole-MoiYoi (Kenya)
Dr Alberto Piazza (Italy)
Dr Svante Pääbo (Germany)
Dr Marcello Siniscalco (Italy)
Dr Ken Weiss (USA)

The HGD Executive Committee, which will have the general responsibility for overseeing the development and implementation of the HGD Project, will meet twice a year.

The HGD Executive Committee will have as its first priorities the identification of a standard set of markers, the development of database software, the establishment of cell- and DNA-banking facilities, the implementation of the pilot project and the development of an information brochure outlining for the public the mission of the project.

Operating rules are being prepared for the **HGD Executive Committee, which will have two standing sub-committees: one concerned with databases and informatics** - initially to consist of Drs Tom Marr, USA (Chairman), Takashi Gojobori, Japan and Otto Ritter, Germany - **and one concerned with ethics.** It was agreed that the latter, which has yet to be constituted, should be subject to the overall guidance of HUGO's standing committee on Ethics, currently co-chaired by Drs Nancy Wexler (USA) and Alain Pompidou (France) and should include members specifically qualified to address the special ethical problems that will be entailed by a world-wide, intercultural endeavour and that may differ from those arising in Western societies or medical clinics. **No other standing sub-committees of the HGD Executive Committee are considered neccessary at the present time.**

3 **At its meeting in January 1994, the Council of HUGO appointed three of its members (Walter Bodmer, UK; Victor McKusick , USA; and Sergio Pena, Brazil) to oversee the functioning of the HGD Executive Committee and to authorise, on behalf of the Council, any public statements or documents about the HGD Project issued by the Executive Committee.**

4 **To encourage and coordinate local involvement in the HGD Project in all regions of the world, regional HGD committees should be established.** It is proposed that a division of the world into about ten or twelve regions would give the necessary coverage. It is recognised that regional affiliations might be ambiguous for some countries but that such problems could be overcome by the Executive Committee in collaboration with the regional committees. Funding mechanisms, for example, could help to decide regional affiliation.

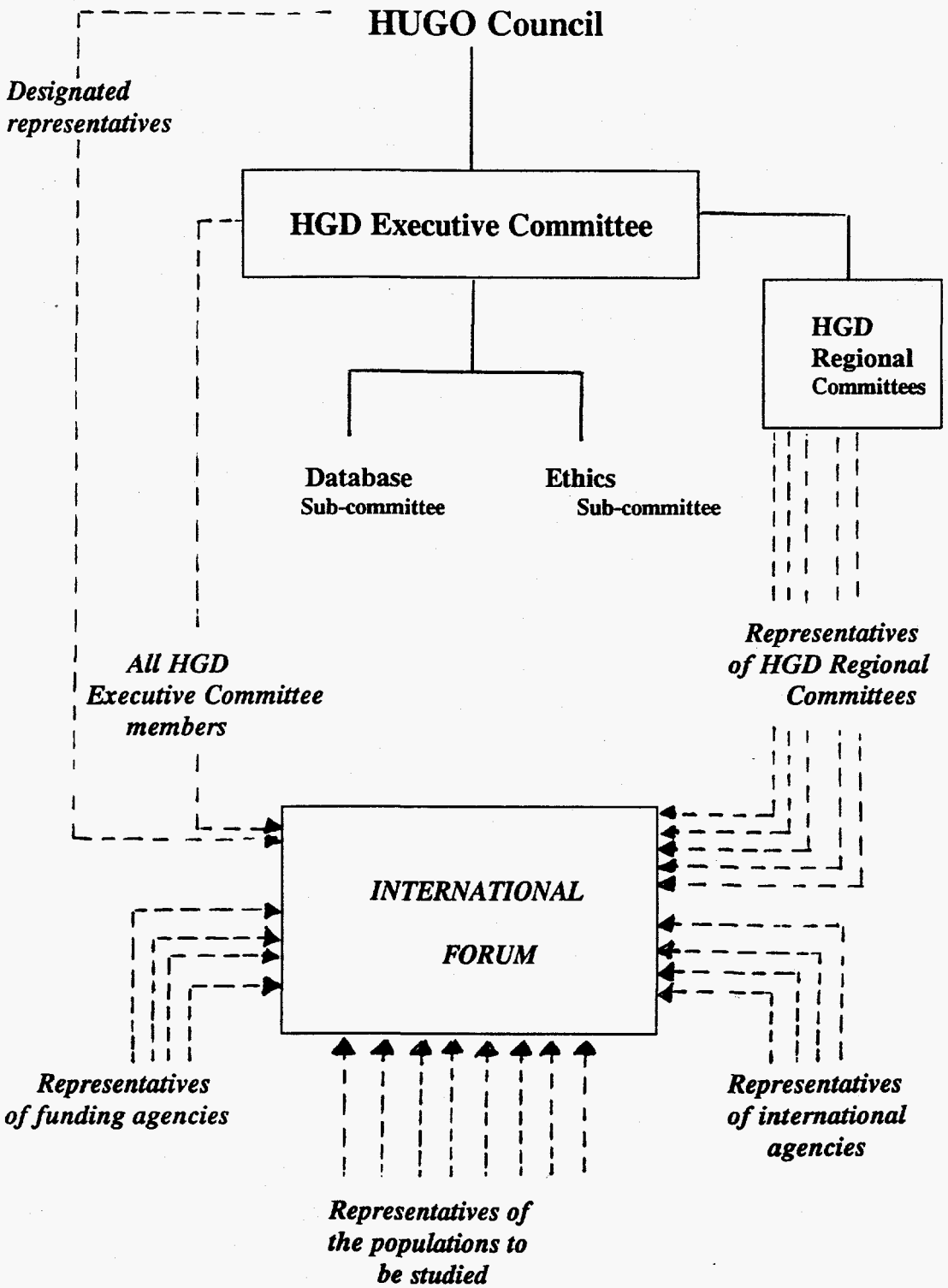
Rather than forming regional committees, the HGD Executive should act to encourage their formation by local investigators and have the power to ratify them as participants in the HGD Project. A regional committee will only be approved by the Executive Committee if it can guarantee that:

- it will conduct and/or approve the collection of samples under ethical standards acceptable to the Executive Committee,
- that it will ensure proper transfer of materials and data to the central repositories, and
- its operation is compatible with that of other regional committees.

The regional committees would seek their own funding, although the Executive Committee would offer its services in this regard where needed and appropriate.

It was recommended that each regional committee consider appointing a subcommittee on local ethical issues. This followed from extensive discussion of regional differences in what consitutes ethical problems in such issues as informed consent. Other functions of the regional committees should be left to the discretion of each regional committee, which would have general responsibility for organising the collection of samples within the region.

Since the workshop, regional committees for North America and for Europe have been formed and the formation of other regional committees is under discussion.


HUGO Council
Designated representatives
HGD Executive Committee
HGD Regional Committees
Database Sub-committee
Ethics Sub-committee
All HGD Executive Committee members
Representatives of HGD Regional Committees
INTERNATIONAL FORUM
Representatives of funding agencies
Representatives of international agencies
Representatives of the populations to be studied

5 **Once a year, the HGD Executive Committee will convene an International Forum,** comprising:

a) Representatives of the regional committees from around the world
b) Representatives of international organisations or bodies with an obvious interest in the HGD Project (eg. UNESCO, WHO) and representatives of the populations to be studied (to be identified by the regional committees).
c) Representatives of funding agencies
d) The HGD Executive Committee
e) The members of HUGO's Council designated to liaise between the Council and Executive Committee

The International Forum will provide an annual opportunity for all the major groups involved in the HGD Project to assess progress, plan for future activities, consider any areas of difficulties, etc.

The International Forum is not intended to be an open meeting but it was agreed that open meetings at which the goals and progress of the HGD Project are reported to the public are desirable and that they could be held in conjunction with the annual meetings of the International Forum.
