Technical Summary of Project Protocol &
Informed Consent Form

-The Genographic Project-

CONTEXT

Technical Summary of Project Protocol

- In the interests of transparency, The Genographic Project has sought and been granted special approval from The Social and Behavioral Sciences Institutional Review Board's (SBS IRB) to release a version the project's US protocol to third parties. Traditionally the specific aims and methods of a research study of this nature are strictly confidential to the researcher(s), their institution, and the review board.

- Because the Genographic Project is a global endeavor from National Geographic's Mission Programs, involving collaboration from indigenous groups, the scientific community, the general public and many audiences, we have maintained from inception that we would like to offer the highest level of transparency, and it is in this spirit we are creating a channel to release this document to third parties.

- The Social and Behavioral Sciences Institutional Review Board's purpose is to protect the rights and welfare of human research participants. Because the IRB must ensure that every aspect of the project maintains the highest ethical and legal standards, the Protocol must detail very specifically the **technically** relevant aspects of the project. The language contained in the Protocol is correspondingly detailed so as to ensure no ambiguity or misperception by the Review Board, and the reader should anticipate this specificity of language and level of detail.

- The attached documents are exactly the same as those approved for the project, other than the removal of the following information:
  Technical Summary of Project Protocol:
  Contact information for the Principal Investigators
  Specific financial budgets for the regional centers
  Database architecture & security, as it is proprietary.
  Any references to Appendices, as these contain information about Principal Investigator CVs, laboratory personnel.
  Consent Form:
  Dr Schurr's contact information.

  The modifications (appropriating the IRB title 'Technical Summary') are so to protect only proprietary information that is not appropriate for public release.

- This (approved) Protocol (required prior to sampling any volunteers from North America for the project), details all aspects of the project that

relate directly to the methods, goals, and procedures relating to the involvement of the human participants. These include the field sampling recruitment and informed consent process, the laboratory methods and research goals, and security and storage of all data and genetic material, as well as scientific collaborators and oversight that will be maintained throughout the project.

- The IRB at University of Pennsylvania is the oversight committee for all research conducted in North America, comprising Canada, USA, Mexico, the Caribbean and Central America. For research conducted outside of North America, each regional research center must also receive approval from their local IRB or ethics review board prior to any field sampling for the project. These regional IRBs are fluent in the local restrictions, consent process, and sensitivities in their region for research involving human participants. All protocols and consent forms that are approved in each regional center will in turn be submitted to the IRB at University of Pennsylvania, where they will also be reviewed to comprise an embedded level of ethical scrutiny.

Informed Consent Form

- The recruitment process for volunteer participants is detailed in the protocol. An important part of this process is ensuring that informed consent is properly obtained prior to the participation of any individual in the research project. The Informed Consent Form is a document that will be provided to each participant to explain the project, its goals and methods, its costs and benefits, and ways for the participant to maintain contact with the project should he or she wish to obtain further information or be removed from the project. In instances where the participant cannot read, or cannot read English, either a translator will be provided or the consent form will be read to them by a leader from their group. The specific language used in this form is such as to ensure that any level of education or experience will be equally able to understand all aspects of the project. The language is also constructed such that any individual who cannot read will also be able to obtain just as much information and understanding of the project by having the consent form read to them.

Technical Summary of Project Protocol

Approved for Release by University of Pennsylvania Social and Behavioral Sciences Institutional Review Board, Office of Regulatory Affairs

I. INTRODUCTION & OBJECTIVES

Introduction and Purpose:

Over the past fifty years, genetics has changed our understanding of human origins. The early work of researchers such as the Hirszfelds (1919), Arthur Mourant (1954), Luca Cavalli-Sforza (Cavalli-Sforza et al. 1994) and others on so-called 'classical' polymorphisms, which paved the way for discoveries such as 'Mitochondrial Eve' in the late 1980s (Cann et al. 1987, Vigilant et al. 1991) and 'Y-chromosome Adam' in the '90s (Hammer 1995, Underhill et al. 2000), allowed genetics to define its role in a field of research traditionally dominated by paleoanthropology and archaeology. Initial skepticism by the anthropological community has been replaced by acceptance of the validity of genetic research. Today, it would be difficult to discuss the origin and spread of our species without drawing on the genetic data produced by laboratories around the world.

However, many gaps remain in our knowledge. At the moment we lack sufficient data to be able to answer some of the most fundamental questions in the field of genetic anthropology. This is particularly true for more recent events in human evolutionary history – those that have occurred over the past 10,000 years. Recent work has focused on specific regions to great success (e.g., Wells et al. 2001; Zerjal et al. 2003; Hurles et al. 2003), but these studies are rare. To trace recent migrations systematically, we will need to pool data on the same set of genetic markers from tens of thousands of samples representing global population diversity, a huge research effort that is virtually impossible to envision with current federal funding priorities.

In a unique partnership, The National Geographic Society, IBM and the Waitt Family Foundation intend to undertake this research effort. Other foundations and individual donors may enter this partnership, and recruitment of these potential partners is ongoing. Our goal is to create a 'virtual biobank' of DNA samples collected from indigenous populations around the world, increasing the number of samples available for study by an order of magnitude. These samples will allow us to address many key anthropological and historical questions.

The Genographic Project, commencing in the first half of 2005 and continuing for five years, will collect DNA samples and data from approximately 100,000 individuals around the world. These individuals will be chosen primarily from indigenous populations, many of which are now in danger of cultural extinction as their members move to cosmopolitan cities and lose touch with the ways of life followed by their ancestors. When this happens, the geographic and cultural context in which their

genetic diversity arose will be lost, thereby devaluing the anthropological utility of the genetic data. At the end of the five years, the Genographic Project will have created a public-domain database of human genetic patterns that can be used by generations of scientists, and will serve as an important resource for educating the public and indigenous communities about the history of our species.

The age of globalization and the Internet have brought the world closer together, but that closeness and the pressures of rapid industrial development also threaten indigenous populations, where numbers are fading, languages are dying out at alarming rates, and oral traditions and customs are lost as the last generations pass on. With the accelerating pace of world development, and the concomitant loss of indigenous cultures around the globe, we may have only a single generation remaining to capture a genetic 'snapshot' of the history of our species before it is lost forever. The proposed studies will yield valuable insights into our past while explicitly underscoring the need to celebrate, and not disenfranchise, the distinct people and cultures that share this world.

Salvage – anthropologist

Background:

The Genographic Project is a major international effort to collect population genetic data from over 100,000 individuals around the world. We will sample those indigenous human populations that retain the clearest context for the genetic patterns – who ideally have lived in the place where they live now, with minimal admixture from surrounding populations. The ultimate goal is to provide an answer to the question 'Where do we come from?' One of the most basic human questions, it can actually be seen as two related queries. The first concerns our origins – where did we originate as a species, and when? The second is about our journey – how did we come to populate the entire world? Genetic research has now answered the origins question resoundingly – we all ultimately trace back to Africa within the past 60,000 years. We even know quite a bit about the great Paleolithic migrations that led to our settling the world's other continents. We still need to discover many of the details of this journey, though, and we actually know very little about the 'black box' that separates the end of the ice age from recorded history. Genographic aims to discover these details.

Each of 10 regional centers will be focused on its own localized sampling and research agenda, with the overarching goal of reaching as wide a range of populations as possible. The list of particular research questions includes the following:

- How many waves of migration were there into the Americas, and was one of these along the coast?
- Could Europeans have migrated to the Americas thousands of years ago?
- Is there a genetic signal from the expansion of indigenous American agriculture – i.e. was it farmers or the culture that moved?
- Did the Inca Empire have a genetic impact on northwestern South America?
- Were there any migrations to South America from the Pacific?

- How do we account for the extraordinary linguistic diversity found in South America – have populations there been separate for a long period of time?
- Can we find genetic signals of now extinct groups (e.g. the Arawaks in the Caribbean) in today's admixed populations?
- Was there any admixture between modern humans and Neanderthals during the European Upper Paleolithic?
- Where did Indo-European speakers originate, and what languages were being spoken prior to their spread in Europe and Asia?
- Did the Celtic expansion of the mid-first millennium BC leave a genetic trail?
- What role did the Uralic speakers play in creating the genetic tapestry of modern Russia and Siberia?
- When did modern humans first colonize the arctic?
- Who were the first inhabitants of the Caucasus region, and why is there such incredible linguistic diversity there?
- Did important imperial conquests have an impact on the genetic landscape of the conquered regions (e.g. did Alexander the Great's armies leave a genetic trail)?
- What role did the Silk Road play in dispersing genetic lineages?
- Where did the Afro-Asiatic languages (including Arabic and Hebrew) originate?
- Can we find genetic traces of the great Middle Eastern empires – Phoenicians, Hittites, etc?
- Who were the aboriginal inhabitants of North Africa, and are the Berbers their direct descendants?
- How much genetic exchange has there been across the Sahara?
- Which African populations harbor the most ancient genetic lineages, perhaps suggesting a geographic origin for modern humans?
- How has European colonialism had an impact on the genetic patterns in Africa?
- Can we trace the origins of the Bantu people – and their expansion across Africa – from the genetic patterns?
- Was there a separate domestication of cattle in Africa, and did this lead to a population expansion?
- Where did the Dravidian speakers originate? Were they the 'first' Indians?
- What role has the Indian caste system had in determining patterns of genetic admixture?
- To whom are the Andamanese most closely related?
- Where did the Han Chinese speakers originate?
- How has the geography of China molded genetic patterns there?
- Who were the aboriginal inhabitants of Indonesia, and was there much genetic exchange with Australia?
- Was there any admixture with *Homo erectus* as modern humans spread throughout Southeast Asia?
- What are the patterns of genetic variation in New Guinea, and do they parallel the extraordinary linguistic diversity there?

- How do the genetic patterns in Australia correlate with the Aboriginal song lines – their own oral histories?
- Can we use genetics to trace the spread of the Polynesians and Micronesians from island to island in the Pacific?
- What impact did migratory bottlenecks/colonialism/disease/ etc. have on the genetic patterns in the Pacific?
- Is it possible to obtain intact DNA from the remains of *Homo erectus* and other extinct hominid species?
- Can we study archaeological material in large enough sample sizes to infer something about population movements? For instance, is there a genetic discontinuity in a particular region that would suggest population replacements during large-scale cultural shifts (e.g. the dawn of the Neolithic, or spread of Indo-European languages)?
- Can ancient remains from animals help to trace the spread of domestication?

The other main focus will be to answer questions about the role of human culture in molding patterns of genetic diversity. Are we truly 'self-made' at the DNA level? And if we share a recent common ancestry, why do we look so different from each other – how was human phenotypic diversity generated? At the moment, we know virtually nothing about this topic. This will be investigated in the second phase of the project, and a separate application to the IRB will be made at that time.

While there have been some small-scale studies of these topics in the past, the sampling to date has simply not been extensive enough, nor have the genetic tools been refined enough, to answer the questions properly. The Genographic Project, by focusing a large budget (by anthropological standards) on these issues, aims to add substantially to our knowledge of human ancestry, migration, and history.

Study Aims:

The project objectives are both scientific and educational. The project will create a global database of our human genetic history with the world's largest database of genetic markers devoted to the study of human origins and migration. We hope to raise awareness among the global community of the threat to indigenous cultures and populations through unique story-telling. Our efforts will help to build scientific capacity in developing countries, as we intend to work with local laboratories and scientific organizations. Finally, in disseminating our findings, we will develop a unique educational project that connects individuals in the global community to our findings from around the globe.

To achieve these ends, the Genographic Project has two components: Research and Education.

The Research component is the focus of this application. It involves collecting a relatively small amount of oral history data together with biological samples from

people in ten geographically distinct locations around the globe. Dr. Schurr is responsible for the North American site, which will act as the coordinating center for the project. There will also be research sites in South America; Western/Central Europe; Middle East/North Africa; Former Soviet Union; Australasia/Pacific; East/Southeast Asia; Sub-Saharan Africa; India; plus analyses of Ancient DNA made available through the Wellcome Trust Ancient Biomolecule Center, Oxford.

The Education component of the project, which has been submitted separately for IRB review, will invite individuals around the world to participate in understanding the objectives and results of the Genographic research. In addition to an interactive web site for mapping and understanding the results of the Genographic research, this phase will offer a personal participation component. The methodological detail for this phase (including a sophisticated approach to assure the anonymity of individuals who choose to participate) is being presented to the IRB at this time. As currently contemplated, it will involve permitting individuals to submit a buccal swab sample for genetic analysis on an anonymous basis. Apart from the processing lab, only the participant will have the numerical key linking him or her to the sample sent for analysis. The samples will be analyzed for the presence of the genetic markers studied in the Genographic research effort. By entering the unique code into the Genographic website, the individual's genotype will be displayed with an accompanying multimedia interpretation.

II. STUDY DESIGN

SAMPLING

Scientists based at ten research centers around the globe will obtain blood samples from indigenous populations living in their geographic region. The Regional Centers will be based at the following locations:

Genographic Regional Centers

North America:
University of Pennsylvania
Theodore G. Schurr
Assistant Professor
Department of Anthropology
University of Pennsylvania
325 University Museum
3260 South Street
Philadelphia, PA. 19104-6398

South America:
Universidade Federal de Minas Girais, Brazil
Fabricio R. Santos
Associate Professor, Ph.D.

Departamento de Biologia Geral, Instituto de Ciências Biológicas,
Universidade Federal de Minas Gerais, Brazil
Av. Antonio Carlos, 6627 C.P. 486
31270-010 Belo Horizonte, MG - Brazil

Western/Central Europe:
Institut Pasteur, Paris (mtDNA)
Lluis Quintana-Murci
INSTITUT PASTEUR
CNRS FRE 2849
Unit of Molecular Prevention and Therapy of Human Diseases
25 rue Docteur Roux,
75724 Paris Cedex 15
France.

Western/Central Europe:
Wellcome Trust Sanger Centre, Cambridge (Y-chromosome)
Chris Tyler-Smith (Dr.)
The Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus
Hinxton
Cambs CB10 1SA
UK

Middle East/North Africa:
American University of Beirut, Lebanon
Pierre A. Zalloua
Assistant Professor
American University of Beirut Medical Center
Department of Internal Medicine and Ob/Gyn
P.O. Box 113-6044
Riad El-Solh, Beirut Lebanon

Former Soviet Union:
Moscow Centre for Medical Genetics, Russian Academy of Medical Sciences
Oleg Balanovsky, PhD
Research Centre for Medical Genetics,
Russian Academy of Medical Sciences
Laboratory of ecological genetics
Moskvorechie 1,
115478 Moscow, Russia

Australasia/Pacific:
LaTrobe University, Melbourne
John Mitchell
Department of Genetics, School of Molecular Sciences

La Trobe University
Melbourne, Victoria, Australia, 3086

East/Southeast Asia:
Fudan University, Shanghai and University of Cincinnati
Li Jin (Dr.)
Fudan University, Shanghai and University of Cincinnati
Cheung Kong Lecture Professor
Fudan University
220 Handan Rd., Shanghai, CHINA

Sub-Saharan Africa:
University of Witwatersrand, Johannesburg
Himla Soodyall (Prof.)
Director: MRC/NHLS/Wits Human Genomic Diversity and Disease Research Unit (HGDDRU)
Division of Human Genetics
National Health Laboratory Service (NHLS)
PO Box 1038, JHB, 2000

India:
Madurai Kamaraj University, Madurai
Ramasamy Pitchappan
Senior Professor and Head
Department of Immunology
School of Biological Sciences
Centre for Excellence in Genomic Sciences
Madurai Kamaraj University
Madurai 625021 Tamil Nadu, India

Ancient DNA: University of Adelaide, Australia
Alan Cooper
Professor
Division of Earth and Environmental Sciences
University of Adelaide
SA5005, Australia

Local IRBs will be responsible for the review and approval of the study. All local laws concerning genetic sampling and anthropological fieldwork will be followed, and any necessary permits will be obtained by the regional research centers. Individuals will be invited to participate by the researchers in the regional centers, often in collaboration with local scientists and others too numerous to be listed here. These would typically include local anthropologists, clinical staff, and group leaders. The regional centers are finalizing their list of personnel, with interviews currently underway at several sites. Complete details and qualifications of all personnel to be

used, including postdocs, technicians, anthropologists, and group leaders will be provided to the IRB once this information is collected from each center.

In each case, a population will be approached through a recognized leader (tribal chief, village leader, council of elders, native corporation, etc.) who will be asked if the group is interested in participating in the project. If the answer is affirmative, then individuals would be asked to participate. Their consent would be given on a hard-copy informed consent form. Many of the regional center principal investigators have submitted informed consent forms that have been used in their regions in the past for study in their regions, and, based on this information, the consent form being used in North America will need to be modified to make it appropriate for each center to reflect issues particular to each region and to satisfy their local ethics review requirements.

Before obtaining any blood or other samples, we will ask participants basic questions about their ancestry and duration of residence in the region. No clinical data or information on medical history will be collected from Genographic participants. The nature of these questions may change during the course of the project as specific questions arise, but, during the first phase of the project, we intend to ascertain the following from each person sampled:

1. Name
2. Age
3. Sex
4. "Mother tongue"
5. Ethnicity
6. Place of birth
7. Mother's native (family) name
8. Mother's native tongue
9. Mother's ethnicity
10. Mother's place of birth
11. Father's native (family) name
12. Father's native tongue
13. Father's ethnicity
14. Father's place of birth
15. Maternal grandmother's and paternal grandfather's ethnicity, if known
16. Maternal grandmother's and paternal grandfather's language, if known
17. Maternal grandmother's and paternal grandfather's places of birth, if known

Blood and Tissue Samples

Genetic studies will be primarily by collection of blood samples. In this procedure, 5-10 ml of blood is taken by venipuncture, carried out by a trained phlebotomist. The blood is then processed in a local laboratory to isolate the white blood cells, which are then lysed, preserving the DNA in a high-EDTA buffer. When the DNA samples are returned to the laboratory at the regional center, the DNA will be extracted, purified

and preserved at –70°C in a locked freezer. These samples will form the core of the Genographic sample collection. In the rare cases where it may be impossible to obtain blood samples for logistical or cultural reasons, buccal swabs will be collected instead. These yield far less DNA, and constitute a less valuable resource, but they can still be analyzed for many genetic markers and are preferable to not sampling at all. DNA obtained from the buccal swabs will only be studied for anthropological purposes, and will undergo the exact same laboratory analysis as DNA from the blood samples.

For ethical, logistical and financial reasons, no transformed cell lines will be established from any of the blood collected in the course of the Genographic Project. Many members of indigenous communities find this process 'unnatural', and are averse to the notion that a sample of their tissue will continue to live on after their deaths. In addition, the time and expense of making cell lines is fairly substantial, and we want to ensure that all of our field research funds go toward collecting and genotyping samples. With the amount of blood to be collected (5-10 ml), we should be able to obtain 500-1000 ug of DNA, which is a significant amount of material for storage and future use. Finally, valid concerns have been raised about the chromosomal rearrangements that take place during long-term passage of cell lines and their effect on the cell's genetic integrity. For all of these reasons, we will only be collecting and storing DNA samples.

We anticipate that the time required to enroll research participants, explain the project, obtain informed consent, ask and answer the questions shown above, and take the blood or cheek swab sample will take no more than 20 minutes per person. For remote populations, it is anticipated that the investigator and not the participants will do the traveling. In some instances, it will be necessary to have translators present in order to detail the project, explain the informed consent form and questionnaire, and to relay questions between the indigenous people and the project scientists that are present. The regional centers will be responsible for arranging translators to be used during sampling, as well as for determining when it will be necessary to provide a translator based on the population that is to be sampled.

Each person sampled will be assigned a randomly-generated alphanumeric code (the Genographic Participant Identification number, or GPID) which they can use to retrieve their results and accompanying explanation from a public website. Along with the actual data generated from the laboratory analysis of anthropological markers, the participant will be presented information pertaining to the migratory history of his genetic ancestry, as well as cultural and historical events that have accompanied the movements of peoples throughout human history. This process entails using a computer to view a public website, and entering the individual's GPID where indicated. Results are displayed directly from the website, and require no mail or additional cost to the participant. In instances where it is appropriate, the information necessary to retrieve this information from the website (the website URL, brief instructions) will be left with an appropriate group leader or local representative. If the individuals do not have access to the internet and they request to be informed of their results, these

results will be mailed or otherwise communicated to the individual by appropriate means. The link between the data and the individual will be destroyed by the regional center after this is done to insure anonymity. The code will be entered into the database, but the person's name will not. In this way the database will be completely anonymous. The field collection sheets, with names and other personal identifiers, will be kept under lock and key at the respective regional center, along with signed consent forms.

Sample Size Determination:

It is common practice in human population genetics research to sample a minimum of 50 individuals per population in order to detect differences in allele frequency between populations. These sample sizes are generally agreed to be the minimum necessary for accurate binomial estimates of marker or haplotype frequencies (Cavalli-Sforza et al. 1994). Of course, this is based on the assumption that each locus is diploid, as is true for autosomal loci. In the case of the Genographic Project, we will initially be examining Y-chromosomal and mtDNA variation, each of which is haploid – i.e., it is only present in a single copy in each individual. Thus, we need to increase our sample sizes to approximately 100 individuals per population to obtain accurate frequency estimates. It is likely, however, that some isolated populations will not be large enough to provide 100 unrelated male donors. In these cases, we will collect as many samples as possible, since, even with small sample sizes, we can often infer phylogeographic information---the geographic origin of markers and their ages relative to other markers in the phylogenetic tree. Thus, our goal will be 100 unrelated samples per population, except in cases where this is impossible. The individuals must be unrelated to avoid re-sampling lineages and thereby overestimating their frequency in the population.

Participants have the option of withdrawing from the study at any time after providing a blood or buccal cell sample. Based on past experience, withdrawal rates are expected to be minimal (< 1%), and should not affect the sampling strategy.

During the course of the five-year project, each research center will collect ~100 individuals from ~100 populations, yielding a total of approximately 10,000 samples per center. Analysis of the samples will take place at the regional centers themselves, and the samples will remain stored in perpetuity at the regional center responsible for their collection, unless the participant chooses to opt out of the project (see below). While the exact number of samples to be collected will vary from site to site, we do not anticipate any regional center collecting more than 20,000 samples during the course of the five-year project.

LABORATORY ANALYSIS

Laboratory analysis of the samples will be focused on identifying genetic markers to infer genealogical relationships. No markers of known medical relevance will be investigated. Our goal is to assess genetic components of ancestry and to discern

their relationship to historical migration patterns. The markers to be typed in the first phase of the project will be Y-chromosome short tandem repeats (STR) and single nucleotide polymorphisms (SNPs), as well as mitochondrial DNA (mtDNA) sequences from hypervariable segement I (HVS-I) and haplogroup-defining non-HVS SNPs. As we will not be sampling related individuals, at least through their maternal and paternal grandparents, the use of the term genealogical relationships is used here in the context of deeper ancestry – i.e., ancestors shared hundreds or thousands of years ago.

Standard PCR and fluorescence-based methods will be used for the genotyping, as described in the literature (Brent et al. 2003). Raw data will be interpreted at the regional centers and prepared for database submission. Both the raw and interpreted data will be submitted to the central Genographic database, to reside at National Geographic Headquarters in Washington, DC. Only anonymized data will be entered into this database, and no one outside of the individual regional centers will have access to the codes allowing names and genotype data to be connected, as described below.

III. SUBJECT SELECTION & WITHDRAWAL

Population sampling will be carried out by the regional centers, assisted in some instances by Dr. Wells. Each center will be responsible for sampling approximately 10,000 individuals from indigenous groups within their region, yielding ~100,000 genotyped samples over the course of the project. In addition to the published data already available, this will yield a database of well over 100,000 individuals typed for a core set of polymorphic DNA markers.

Individuals from indigenous groups will be chosen because they have a lower likelihood of extensive recent admixture with other population groups. While many of the markers we plan to study could be examined in cosmopolitan populations, the geographic and cultural context of the data would be minimal. We will use the data to infer historical patterns of migration and demography, and hope to mitigate the effects of recent migration by choosing 'isolated' groups. While no population has been completely isolated throughout its history, we hope to bias in favor of this by choosing primarily non-cosmopolitan groups for sampling.

Only men over the age of 18 will be sampled. The age of 18 is generally agreed to be the international age of consent. Very importantly, men, as bearers of both mtDNA (inherited from their mother) and a Y-chromosome (from their father), maximize the amount of useful genetic information that can be obtained from the samples. If we were to sample equal numbers of men and women we would only obtain half the number of Y-chromosomes. In addition, because men have one X and one Y chromosome, rather than women's two X chromosomes, it is easier to study patterns of X chromosome variation in men if we decide to do so in the future. This is because haplotypes (linked sets of genetic markers that are useful for inferring migratory

history) are easily defined for single-copy chromosomes, but this process is much more difficult when two copies are present. Thus, there are sound scientific reasons for not sampling women. Women will of course be equally represented in all of the ethnographic information we collect – their contribution to the population is equally important – and in the mtDNA and autosomal data collected from men (in effect, by sampling them we are sampling their mothers as well). The reasons for not taking samples from women are purely technical.

The remaining inclusion criteria are straightforward: belonging to one of the hundreds of groups that will yield informative results for our studies of human migration and 'deep ancestry'. To be included in the sampled group, the participant will need to have grandparents who were members of the population in question. This is to help account for recent admixture in the population, and to assure that the genetic lineages we find are as representative of the ancient history of the population as possible.

The choice of populations will take place in two ways. Some populations may contact the Genographic Project asking to be included in the study. This has happened to Dr. Wells and his colleagues on many occasions, and we anticipate this being a potentially important component of the Genographic project. Alternatively, members of the Genographic scientific team will choose populations to approach for inclusion on the basis of sound scientific or historical criteria – *i.e.*, to answer specific scientific questions. In this case, recruitment for the Genographic Project generally will involve establishing relationships with individuals of a given community (e.g., elder tribe members, village council members, members of a particular indigenous group, etc.) and visiting remote areas where these communities are located to identify individuals who would be interested in participating in the Project. Where travel is required to meet with the indigenous group being sampled, the research team will travel to the location of the indigenous people. Once at the site, the researchers will meet individually with all interested participants and read the consent form, which detail the purpose and process of collecting the samples (see attached). In cases where a translator is necessary but a member of the group cannot translate the consent form to the group members, a translator will be provided by the regional center PI.

In both cases, though, recruitment will be carried out in such a way as to reassure the participants that they are true collaborators in the scientific process and not merely 'study subjects'. Extensive notes will be compiled of group ethnographic data that may be added to the Genographic database at a future time – marriage patterns, demographic events, oral history, etc. In the event that this information is collected, it will initially be recorded by the researcher on location. These data will subsequently be transferred to the central database at National Geographic headquarters, where they will be stored for reference by future analyses. Once the data has been submitted to the central database, the record from the field researcher (either on paper or on laptop computer) will be destroyed or deleted permanently. In the case of fragmented populations belonging to the same self-described ethnic group, an effort will be made to sample across several groups in order to asses both intra- and inter-group heterogeneity.

We will not sample from prisoners, seriously ill individuals, captive populations, or other vulnerable groups. Only healthy adult males will be asked to participate. Individuals who are hospitalized or under physician care for an acute condition will not be included for this study. Individuals who have a chronic but non life-threatening condition may be included in the study; however, in this and any other case, no medically relevant information will be recorded for any individual.

Due to the nature of the project – its global scope, multiplicity of languages, and preliminary stage of development (e.g., not having identified all populations to be sampled) – it is impossible to provide examples of the recruitment materials for non-U.S. sites at this time. We will forward additional materials to the IRB as they are created, and hope to gain approval at this stage for the aims and general structure of the project, as well as for the consent forms and data collection plans for the U.S. site(s).

Inclusion Criteria:

See above.

Exclusion Criteria:

- <18 years old
- Female
- Prisoner
- Psychiatric illness interfering with cognitive function
- Dementia/mental retardation
- Serious/fatal illness

Subject Recruitment and Screening:

This was discussed above, and will vary according to region and population. All recruitment materials for non-U.S. sites are being submitted to the IRB for approval as they are created. Both of these methods will be explained to the participants during sample collection, and contact information for the regional center will be left with them by the research team (as described above).

Early Withdrawal of Participants:

Participants can withdraw from the study at any point and for any reason in two ways: by contacting the regional center by telephone or in writing, giving their name and assigned code number, or by opting out of the project via the secure website (as described above). The appropriate contact information for each site will be established once all personnel are finalized for each site and before sampling has begun.

IV. METHODS AND INSTRUMENTS

Study Instruments:

See above for questions to be asked in the interview process.

In some cases, participants or other individuals may be approached for inclusion in a documentary film, magazine article or book intended for broadcast or publication. In these cases specific informed consent will be obtained to cover these activities, as detailed in the attached consent form.

Group modifications:

At this time, due to the preliminary nature of the project, it is difficult to address this issue. If group modifications are necessary, then we will submit protocol amendments.

Maintenance of data security:

Local/Regional Database. Each of the ten sites will develop procedures for ensuring that (1) the names and any other identifying information, such as social security or ID numbers, addresses, and the like, are not retained or transmitted as part of the data for the project, and (2) the consent and any other information retained at the site do not include the GPID of the participant.

Site-Specific Security during Data Collection. Each site will obtain and analyze the biological materials collected from participants in accordance with the instructions of the local review board and the consent of the individual. The results of the genetic analysis of the sample will be recorded under the GPID for transmission to the database. Unless required to do otherwise by the local review and/or ethics boards, we will retain the identity of the participants on only the consent form to participate in the Genographic Project. These forms, and any data linking them to the GPID, will be kept under lock and key at the research center for that respective region. Only the Principal Investigator and postdoc from the center will have access to these forms.

The Genographic data are proprietary to the Genographic Project. The data will be maintained in computer databases, even at the local sites, to the extent possible. We will require each site to maintain strict security of the computer by using individual biometric IDs, login IDs and passwords to restrict access to the proprietary databases, and to keep the computers locked away from personnel not authorized to work on the project. We will require these databases to be purged at the conclusion of the site's involvement in the Genographic Project.

Genographic Database. In the data to be maintained for analysis in the Genographic database at National Geographic headquarters in Washington, DC, participants from each of the ten sites will be identified only with a "Genographic Participant Identification number" (GPID) that records the site and accession number of each case. Gender, age, and some ancestry information will be maintained for each, but no medical, health or disease information will be solicited or recorded for the Genographic Project. Likewise, the genetic analyses of the sample obtained from each participant will be recorded under the same GPID for inclusion in the database.

National Geographic has no intention of selling the data or patenting any discoveries resulting from the project. All research data will be released into the public domain at the time of its publication in scientific journals. Because the Genographic Project is a partnership between National Geographic and IBM, one party is required to "own" the data, and it is National Geographic that will take on this role.

See below for additional information on the database security procedures.

V. STUDY PROCEDURES

The sample and any oral history information will be collected from each participant in a single visit as described above.

VI. STATISTICAL PLAN

Statistical Methods:

Many statistical methods will be applied to the analysis of the data generated during the Genographic project, including new methods that are likely to be developed during the course of the project. The goal is to detect historical migration events by inferring the origin and spread of genetic lineages (phylogeography). The main analyses can be divided into three categories:

- Population-based methods based on allele frequencies. Phylogenetic methods using genetic distances can be useful in summarizing population relationships, as can principal components analyses (Cavalli-Sforza et al, 1994). Analysis of

molecular variance (Excoffier et al. 1992) allows comparisons of within-versus-between patterns of genetic variation. These methods compare discrete population units to each other in an effort to discern which populations are more closely related.

- Geographic methods based on allele or haplotype frequencies, such as spatial autocorrelation (Sokal and Oden 1978). These methods are used to test for uneven geographic distributions of genetic markers that may be indicative of migration events, and are not limited to comparisons of discrete population units.

- Evolutionary methods based on mutation rate, used to estimate the branching pattern and ages of genetic lineages. Examples include those implemented in NETWORK (Bandelt et al. 1999) and BATWING (Wilson and Balding 1998).

Subject Population(s) for Analysis:

The Genographic project is designed to be carried out on a global scale, and samples will be collected from individuals living in the majority of the world's countries. The regional centers are based in established laboratories known for their previous work in the field of human population genetics. Each field center will enlist a network of local collaborators in its region to assist in carrying out the sampling. The sampling work at each center will be overseen by a Principal Investigator and a postdoctoral fellow. The postdoctoral fellow will carry out much of the field work, assisted by graduate students, technicians and trained phlebotomists. The project has a substantial budget for travel and related collecting expenses to allow the personnel to reach remote populations.

It is impossible at this time to provide a detailed list of the populations that will be asked to participate in the Genographic Project, as these will change during the course of the project. In the future, as populations that might participate in the project are identified we will submit these for approval to the advisory board, and submit copies of local IRB approvals to the U. of Penn IRB.

Data Management:

The National Geographic Society is responsible for maintaining the security of the data and data handling systems. To ensure the highest security possible, we have partnered with IBM to design and develop the data handling and access mechanisms. We will adhere to the highest standards of data security.

VII. ETHICAL CONSIDERATIONS:

Advisory Board:

The Genographic Project has in place an Advisory Board to review its procedures, methods, and arrangements for ensuring ethical behavior in its dealings with indigenous people and more generally in its approach. Advisory Board members are not paid for their services. The Advisory Board will meet regularly during the project to evaluate the work and resolve any issues that have arisen in connection with one or more of the project sites.

The Advisory Board Members include:

Name	Affiliation	Area of Expertise
Dr. Luca Cavalli-Sforza, M.D. Chair	Stanford University	Human Population Genetics
Lord Colin Renfrew, Ph.D.	Cambridge University	Archaeology
Dr. Merritt Ruhlen, Ph.D.	Stanford University	Linguistics
Dr. Meave Leakey, Ph.D.	National Museums of Kenya	Paleontology
Dr. Scott Edwards, Ph.D.	Harvard University	Evolutionary genetics
Dr. Wade Davis, Ph.D.	National Geographic Society	Cultural anthropology
Tammy Williams	Indigenous Enterprise Partnerships, Cape York, Queensland	Indigenous Rights Advocate
Dr. Simon Longstaff	St. James Ethics Centre, Australia	Ethics
Nick Donofrio	IBM	Information technology
Terry Garcia	National Geographic Society	Project management

Anticipated Benefits:

Individual study participants will not be providing any health or other medically relevant information, and the genetic analyses will not provide information that can be used in health or health care planning. The primary benefit to individual participants will come from the knowledge that they will gain about their deep ancestry and relatedness to people in other regions of the world. Our local experts at each site will be developing narrative methods to help indigenous peoples understand the role of their ancestors in human history.

Furthermore, through the Genographic database and public website, we aim to provide a valuable resource for the general public to learn about our common ancestry. In today's fractious world, knowledge of the genetic ties that bind us

together in one large, extended family may help us to mitigate some of the prejudice and hatred that have defined human relations for thousands of years.

Finally, the Genographic database and the virtual biobank of genetic samples from indigenous populations will provide a 'genetic snapshot' of our species at this point in time – a time of great social change, which threatens to destroy much of the world's cultural legacy. Generations of researchers will be able to make use of these resources, perhaps applying research methods that we cannot envision at the present time. It is hoped that this scientific legacy will continue to add to our knowledge of the origin and spread of our species over the past 200,000 years.

Risks:

We are not aware of any medical risks of participating in the Genographic Research project other than the relatively minor risks associated with drawing a small sample of blood. There is minimal risk to the subject associated with blood collection by venipuncture, usually limited to bruising and discomfort at the site of needle entry. Sterile procedures will be followed at all times, so the risk of subject infection is extremely low. The biohazard risk to the investigator is also minimal, and the person carrying out work on the blood samples will be trained to follow procedures to reduce the risk of blood-borne pathogen transmission (due to needle sticks, etc.).

The genetic analyses performed as part of the Genographic Project will not reveal paternity or heritable disease status that might cause emotional distress, create family tensions, or lead to discrimination by healthcare or insurance providers. Moreover, the markers we are examining will not categorize participants as members of a particular race or ethnicity, and will only be used to discern ancestral migratory relationships to other populations. It is possible that an individual's genetic markers may reveal their ancestry to be more complex than they initially suspected, and may contradict some details of their assumed personal history and group affiliation(s), but this possibility will be explained during the consent process. At any point and for any reason, a participant may decide to opt out of the study by simply contacting the research center either by mail, phone or e-mail, at which point all records of data relating to the sample will be destroyed along with the sample itself. Care will be taken to explain the goals and results of the research in a culturally sensitive way. Ultimately, participation is voluntary, and no results can be guaranteed to the participants.

Risk/Benefit Ratio:

The risks to the individual from participation are essentially those of a routine blood draw, while the benefits of the project offer humankind as well as the individual a more complete understanding of the deep ancestry of our species and the way we have populated the globe. Since each individual's participation is entirely voluntary, and will result in no direct benefit to the individual other than knowledge of certain aspects of their ancestry, it is our hope that everyone who elects to participate will be interested

in the benefit of the knowledge obtained, and hope that they feel that it outweighs the individual risk from the routine blood draw.

Subject Confidentiality:

Genographic Participant Number. In the data to be maintained for analysis in the database at National Geographic headquarters in Washington, DC, participants from each of the ten sites will be identified using only a Genographic Participant Identification number (GPID) that records the regional center and accession number of each case. Gender, age, and some ancestry information will be maintained for each, but no medical information will be solicited or recorded. Likewise, the genetic analyses of the sample obtained from each participant will be recorded under the same GPID for transmission to the central database.

No Key to Link GPID to Participant Identity. Each site will obtain and analyze the biological materials collected from the sample donors in accord with the instructions of the local review board and the consent of the individual. The results of the genetic analysis will be recorded under the GPID for transmission to Genographic database. Unless we are required to do otherwise by the local review and/or ethics boards, the identity of the participants will be retained only on the consent form to participate in Genographic Research. All links will be kept under lock and key at the regional center, with access limited to the PI, who will be responsible for their safekeeping.

Local Review Boards. To the extent that a local review board permits or requires investigators at a site to retain any materials or data for purposes unrelated to the Genographic Project, each site will comply with local rules and requirements. However, it will not be permitted to retain the GPID or other data that reasonably could be used to link any locally retained information to the Genographic Research data. The Genographic Project will maintain compliance with the ethical requirements established in each region of study, as determined by the local Institutional Review Board (IRB) or Ethics board that oversees research studies in that region.

Anonymity of Database. For purposes of compliance with the data protection laws of other countries and of the USA, the data transmitted to the database in Washington, DC will be anonymous, as no key or linking code will be in existence. We consider the Genographic data to be proprietary to the Genographic Project, so we will impose strict confidentiality and security requirements on the regional centers. We also will require them to segregate any data created for the Genographic Project from other data they create or maintain.

Once the Genographic data are submitted to the database, they will be accessible in this anonymized form by all researchers involved in the Genographic Project. When the results are published in scientific journals, the relevant data will be made publicly available as required by the journals where the data is published. Individual participants will not be identified by name in any publications or materials generated

from the project, unless fully informed and written consent is first obtained from the participant.

At the completion of the project, the database will be made available in its complete, anonymized form to the public.

No Medical Information. The creation and transmission of the genetic and oral history data are not subject to the medical privacy laws of any country, including the HIPAA regulations in the United States, because they will have been stripped of all personal identifiers. None of the data relates to the health or health care of the individual, even where the genetic analyses are performed by laboratories that may be subject to medical privacy requirements with respect to other parts of their business. Consequently, the data are not protected health information. No biological materials or samples will be transported from any of the non-US sites to the US.

Reimbursement:

Participants will not be compensated except in a modest reimbursement as part of a normal professional interaction – e.g. offering some food or a beverage in the setting of a waiting area, etc. We will, however, reimburse reasonable travel expenses (train tickets, taxi cab fares, child care costs, etc.) for subjects who need to travel to the collecting site. No monetary or other inducement of any kind will be offered to potential participants in order to assure that their participation in the project is truly voluntary.

VII. INFORMED CONSENT

Each person who participates in the Genographic Project will provide informed consent by signing the attached informed consent form before the blood draw and collection of oral history/family information. In addition, consent will be obtained from the group to which the individual belongs if this is necessary within their particular culture.

The consent forms provided for the North American site will be translated into appropriate languages for submission to the local review board and/or authorities.

VIII. FINANCIAL CONSIDERATIONS

Funding Source:

The Genographic Project will be funded by National Geographic, with donations solicited from foundations and individuals. IBM will provide the in-kind support needed

to perform the analyses, as well as the establishment and operation of the Genographic web site.

IX. PUBLICATION PLAN

The preferred method of publishing the results of the Genographic Project will be in the peer-reviewed scientific literature. All publications will list the Genographic Consortium (all of the scientific personnel active in the project) as follows:

The first author will typically be the **postdoc** or other individual from the regional center, IBM or NGS, or the person who did the lion's share of the work,

Followed by

Other **significant contributors** from the regional center, IBM or NGS (technicians, graduate students, major collaborators),

Followed by

The **PI** of the regional center, IBM or NGS,

Followed by

The Genographic Consortium, defined in a footer, to include all PIs and postdocs from the regional centers, the core IBM research team, the National Geographic team, and Spencer Wells.

In addition, as results are published, the media platforms of National Geographic (NG magazine, NG Channel, the NG website) will convey the results to the general public, including the indigenous populations involved in the research. The Genographic website (http://www.nationalgeographic.com/genographic) will be a major conduit for these stories. The results will be displayed using maps, visual images, and narrative to integrate historical, anthropological, linguistic, climatic, and archaeological findings in an effort to provide a comprehensive understanding of the genetic results in the context of human history. There will be areas of the web site that can be used by teachers in classes at all grade levels, as well as areas that individuals can explore at their leisure.

References

Bandelt, H.J., Forster, P. & A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16**:37-48.

Brent, R. et al. 2003. *Current Protocols in Molecular Biology*. John Wiley and Sons.

Cann, R.L., M. Stoneking, M & A.C. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* **325**:31-36.

Cavalli-Sforza, L.L., Menozzi, P. & A. Piazza. 1994. *The History and Geography of Human Genes*. Princeton University Press.

Excoffier, L., Smouse, P.E. & J.M. Quattro. 1992. Analysis of molecular variance inferred from metric differences among mtDNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**:479-491.

Hammer, M.F. 1995. A recent common ancestry for human Y-chromosomes. *Nature* **378**:376-378.

Hirszfeld, L. & H. Hirszfeld. 1919. Serological differences between the blood of different races. The result of researches on the Macedonian front. *Lancet* **ii**: 675-679.

Hurles, M.E. et al. 2003. Native American Y chromosomes in Polynesia: the genetic impact of the Polynesian slave trade. *American Journal of Human Genetics* **72**:1282-1287.

Mourant, A.E. *The Distribution of the Human Blood Groups*. 1954. Oxford: Blackwell Scientific Publications.

Sokal, R.R. & N.L. Oden. 1978. Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society* **10**:199-228.

Underhill, P.A. et al. 2000. Y chromosome sequence variation and the history of populations. *Nature Genetics* **26**:358-361.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & A.C. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. *Science* **253**: 1503-1507.

Wells, R.S. et al. 2001. The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences USA* **98**:10244-10249.

Wilson I.J. and D.J. Balding. 1998. Genealogical inference from microsatellite data. *Genetics* **150**:499-510.

Zerjal, T. et al. 2003. The Genetic legacy of the Mongols. *American Journal of Human Genetics* **72**:717-721.

National Geographic Society
Social & Behavioral Science

Consent Form – North America

Title of Research Project: "*The Genographic Project: Molecular Genetic Analyses of Indigenous Populations of North America – University of Pennsylvania*"
Protocol Number: 803115
Principle Investigator: Dr. Theodore G. Schurr, Department of Anthropology, University of Pennsylvania
Overall Project Principle Investigator: Dr. Spencer Wells, Mission Programs, National Geographic Society, 1145 17th Street, N.W., Washington, D.C. 20036

Invitation to Participate: You are invited to participate in a study of different groups of people who lived long ago in North America. The project team will talk to you about the project and how it might affect you. They will explain what you will have to do if you decide to participate. You can choose whether or not to participate. If you decide not to participate, no one will be angry with you.

Please read this consent form carefully. If you do not understand something, please ask the researcher to explain it. If you prefer, someone will read this to you. If you decide to participate, please sign the last page of this form. We will give you a copy to keep.

Purpose: This project is a major international effort to collect population genetic data from over 100,000 individuals around the world. We are sampling those indigenous human populations that retain the clearest context for the historical events that have contributed to the current genetic patterns – who ideally have lived in the place where they live now, with minimal intermarriage from surrounding populations. Our goal is to understand the relationships between genetic, linguistic, cultural and historical data. No clinical data or information on medical history will be collected from you, and thus your blood will never be used for any medically-related inquiry. To better understand the history of the indigenous peoples from your region and around the world, the investigators will survey segments of your DNA for variation. Small changes in these portions of your DNA create markers that can be used for population studies. These changes are helpful in studying human history because many of them cluster together in specific ancestral lines of descent within human populations. Because of how these portions of DNA are passed from parent to child, we can trace these mutations through human families from the present to the distant past with a relatively high degree of accuracy. We can also reconstruct patterns of human movement through geographic areas by tracking the spread of these lines of descent in different human groups.

Length of the Project: Your participation in the project will take no more than 20 minutes. You will be one of many thousands of indigenous people in your region who

give samples. We will be here for only a short visit to collect the samples, but the entire project will last 5 years.

Procedure: If you decide to participate, we will draw about 5-10 ml (1-2 teaspoons) of blood from a vein in your arm, using a small needle. The blood will be used to study the different groups that contributed to your family. We will also ask questions about your family history.

Storage of Sample: Your blood sample will remain stored at the Laboratory of Molecular Anthropology at the University of Pennsylvania after the completion of this project, where it may be used for further study to better understand human origins and the histories of your people. The blood sample cannot and will not be used for any medically related study. Furthermore, you may at any point in the future choose to have your sample removed from the project. You can do this by simply contacting our regional center and explaining your desire to have your sample removed. Your sample will then be immediately destroyed at the laboratory and all records and associated data from your sample will be eliminated. Instructions on how to do this are explained below.

Risks: You will spend about 20 minutes on the project. When we take the blood sample, you may feel a little pain (like a pin prick). The place where the needle goes into your arm may bruise or hurt a little afterwards. Very rarely, someone will develop an infection where the needle was inserted. You will have no other pain or risk to your health. All blood samples will be drawn by doctors or trained assistants. It is possible that some of the findings that result from this study may contradict an oral, written or other tradition held by you or by members of your group. Ultimately, participation is voluntary and we cannot guarantee any results to you.

Benefits: The only benefits to you from participating in this project will be to learn more about your family origins and your relationships to other people around the world. You will be able to learn this by going to the project website at www.nationalgeographic.com/genographic Once on the website you will enter the secret code assigned to your sample which we will give you here, and you will then be able to view the results of your own family's genetic history. In easy-to-understand language you will be able to learn things about the different regions that your ancestors crossed around the world, and how long ago this journey took place that ultimately brought you here. If you do not have access to the internet, you may contact the laboratory listed below and your results will be mailed to you.

Costs: You do not have to pay anything to be in the project. You will not be paid for being in this project. If you have to travel to the collecting site, we may pay for your reasonable travel costs. We will not pay for your travel costs, *unless* you arrange for them to be paid for before you provide a blood sample.

Participation: The choice is yours. You may choose either to be in the project, or not to be in the project. Your participation is voluntary. There will be no penalty if you choose not to participate or if you agree to be in the project, but later change your mind.

You may withdraw from the project at any time by calling or writing Theodore Schurr at (Tel. #), University of Pennsylvania, 325 University Museum, 3260 South Street, Philadelphia, PA 19104-6398. You may also withdraw from the project through the website at www.nationalgeographic.com/genographic. To withdraw from the project, you must provide your name and assigned code number.

Privacy: We will protect carefully the information that you tell us about yourself and your family. What we learn from your blood sample will be described only in a way that does not identify you. You will be given a code number for your blood sample and instructions how to use the code number to obtain the results of your blood sample. To protect your privacy, blood samples will be recorded with a secret code. Your name only will be recorded on the consent form. The secret code assigned to your blood sample in North America will be kept in a locked file at the site where you participate in the project and carefully protected. No information related to your medical history will be included. Your blood sample will be stored at the regional site unless you ask to have your sample destroyed after the project.

Project Contact(s): If you have any questions about your rights or welfare as a participant in this project, please contact, at the University of Pennsylvania:

- The Social and Behavioral Sciences Institutional Review Board: 215-746-6274; or
- The University of Pennsylvania's Office of Regulatory Affairs: 215-898-00820.

If you have specific questions about the project, please contact Dr. Theodore G. Schurr.

When you sign this page, you are saying that you understand our project, your questions have been answered to your satisfaction, and that you will take part in this project.

Name of Subject	**Signature of Subject**	**Date**
Name of Person Obtaining Consent	**Signature of Person Obtaining Consent**	**Date**