EMBARGOED for Release
Tues., Jan. 22, 2008
8 a.m. Eastern

Contact
Geoff Spencer, NHGRI
301-402-0911
spencerg@mail.nih.gov

International Consortium Announces the 1000 Genomes Project

Major Sequencing Effort Will Produce Most Detailed Map Of Human Genetic Variation to Support Disease Studies

An international research consortium today announced the 1000 Genomes Project, an ambitious effort that will involve sequencing the genomes of at least a thousand people from around the world to create the most detailed and medically useful picture to date of human genetic variation. The project will receive major support from the Wellcome Trust Sanger Institute in Hinxton, England, the Beijing Genomics Institute, Shenzhen (BGI Shenzhen) in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH).

Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

"The 1000 Genomes Project will examine the human genome at a level of detail that no one has done before," said Richard Durbin, Ph.D., of the Wellcome Trust Sanger Institute, who is co-chair of the consortium. "Such a project would have been unthinkable only two years ago. Today, thanks to amazing strides in sequencing technology, bioinformatics and population genomics, it is now within our grasp. So we are moving forward to build a tool that will greatly expand and further accelerate efforts to find more of the genetic factors involved in human health and disease."

Any two humans are more than 99 percent the same at the genetic level. However, it is important to understand the small fraction of genetic material that varies among people because it can help explain individual differences in susceptibility to disease, response to drugs or reaction to environmental factors. Variation in the human genome is organized into local neighborhoods called haplotypes, which are stretches of DNA usually inherited as intact blocks of information.

Recently developed catalogs of human genetic variation, such as the HapMap, have proved valuable in human genetic research. Using the HapMap and related resources, researchers already have discovered more than 100 regions of the genome containing genetic variants that are associated with risk of common human diseases such as diabetes, coronary artery disease, prostate and breast cancer, rheumatoid arthritis, inflammatory bowel disease and age-related macular degeneration.

However, because existing maps are not extremely detailed, researchers often must follow those studies with costly and time-consuming DNA sequencing to help

pinpoint the precise causative variants. The new map would enable researchers to more quickly zero in on disease-related genetic variants, speeding efforts to use genetic information to develop new strategies for diagnosing, treating and preventing common diseases.

The scientific goals of the 1000 Genomes Project are to produce a catalog of variants that are present at 1 percent or greater frequency in the human population across most of the genome, and down to 0.5 percent or lower within genes. This will likely entail sequencing the genomes of at least 1,000 people. These people will be anonymous and will not have any medical information collected on them, because the project is developing a basic resource to provide information on genetic variation. The catalog that is developed will be used by researchers in many future studies of people with particular diseases.

"This new project will increase the sensitivity of disease discovery efforts across the genome five-fold and within gene regions at least 10-fold," said NHGRI Director Francis S. Collins, M.D., Ph.D. "Our existing databases do a reasonably good job of cataloging variations found in at least 10 percent of a population. By harnessing the power of new sequencing technologies and novel computational methods, we hope to give biomedical researchers a genome-wide map of variation down to the 1 percent level. This will change the way we carry out studies of genetic disease."

With current approaches, researchers can search for two types of genetic variants related to disease. The first type is very rare genetic variants that have a severe effect, such as the variants responsible for causing cystic fibrosis and Huntington's disease. To find these rare variants, which typically affect fewer than one in 1,000 people, researchers often must spend years on studies involving affected families. However, most common diseases, such as diabetes and heart disease, are influenced by more common genetic variants. Most of these common variants have weak effects, perhaps increasing risk of a common condition by 25 percent or less. Recently, using a new approach known as a genome-wide association study, researchers have been able to search for these common variants.

"Between these two types of genetic variants — very rare and fairly common — we have a significant gap in our knowledge. The 1000 Genomes Project is designed to fill that gap, which we anticipate will contain many important variants that are relevant to human health and disease," said David Altshuler, M.D., Ph.D., of Massachusetts General Hospital in Boston and the Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard University in Cambridge, Mass., who is the consortium's co-chair and was a leader of the HapMap Consortium.

One use of the new catalog will be to follow up genome-wide association studies. Investigators who find that a part of the genome is associated with a disease will be able to look it up in the catalog, and find almost all variants in that region. They will then be able to conduct functional studies to see whether any of the catalogued variants directly contribute to the disease.

The 1000 Genomes Project builds on the human haplotype map developed by the International HapMap Project. The new map will provide genomic context surrounding the HapMap's genetic variants, giving researchers important clues to

which variants might be causal, including more precise information on where to search for causal variants.

Going a major step beyond the HapMap, the 1000 Genomes Project will map not only the single-letter differences in people's DNA, called single nucleotide polymorphisms (SNPs), but also will produce a high-resolution map of larger differences in genome structure called structural variants. Structural variants are rearrangements, deletions or duplications of segments of the human genome. The importance of these variants has become increasingly clear with surveys completed in the past 18 months that show these differences in genome structure may play a role in susceptibility to certain conditions, such as mental retardation and autism.

In addition to accelerating the search for genetic variants involved in susceptibility to common diseases, the map produced by the 1000 Genomes Project will provide a deeper understanding of human genetic variation and open the door to many other new findings of significance to both medicine and basic human biology.

The sequencing work will be carried out at the Sanger Institute, BGI Shenzhen and NHGRI's Large-Scale Sequencing Network, which includes the Broad Institute of MIT and Harvard; the Washington University Genome Sequencing Center at the Washington University School of Medicine in St. Louis; and the Human Genome Sequencing Center at the Baylor College of Medicine in Houston. The consortium may add other participants over time.

The project depends on large-scale implementation of several new sequencing platforms. Using standard DNA sequencing technologies, the effort would likely cost more than $500 million. However, leaders of the 1000 Genomes Project expect the costs to be far lower – in the range of $30 million to $50 million – because of the project's pioneering efforts to use new sequencing technologies in the most efficient and cost-effective manner.

In the first phase of the 1000 Genomes Project, lasting about a year, researchers will conduct three pilots. The results of the pilots will be used to decide how to most efficiently and cost effectively produce the project's detailed map of human genetic variation.

The first pilot will involve sequencing the genomes of two nuclear families (both parents and an adult child) at deep coverage that averages 20 passes of each genome. This will provide a comprehensive dataset from six people that will help the project figure out how to identify variants using the new sequencing platforms, and serve as a basis for comparison for other parts of the effort.

The second pilot will involve sequencing the genomes of 180 people at low coverage that averages two passes of each genome. This will test the ability to use low-coverage data from new sequencing platforms to identify sequence variants and to put them in their genomic context.

The third pilot will involve sequencing the coding regions, called exons, of about 1,000 genes in about 1,000 people. This is aimed at exploring how best to obtain an

even more detailed catalog in the approximately 2 percent of the genome that is comprised of protein-coding genes.

During its two-year production phase, the 1000 Genomes Project will deliver sequence data at an average rate of about 8.2 billion bases per day, the equivalent of more than two human genomes every 24 hours. The volume of data – and the interpretation of those data – will pose a major challenge for leading experts in the fields of bioinformatics and statistical genetics.

"This project will examine the human genome in a detail that has never been attempted – the scale is immense. At 6 trillion DNA bases, the 1000 Genomes Project will generate 60-fold more sequence data over its three-year course than have been deposited into public DNA databases over the past 25 years," said Gil McVean, Ph.D., of the University of Oxford in England, one of the co-chairs of the consortium's analysis group. "In fact, when up and running at full speed, this project will generate more sequence in two days than was added to public databases for all of the past year."

The 1000 Genomes Project will use samples from volunteer donors who gave informed consent for their DNA to be analyzed and placed in public databases. NHGRI and its partners will follow the extensive and careful ethical procedures established for previous projects. As was the case for the International HapMap Project and Human Genome Project, the 1000 Genomes Project will have an expert working group devoted to examining the ethical, legal and social issues related to its research.

The first thousand samples for the 1000 Genomes Project will come from those used for the HapMap and from additional samples in the extended HapMap set, which used the same collection processes. No medical or personal identifying information was obtained from the donors, and the samples are labeled only by the population from which they were collected. The donors' anonymity was enhanced by recruiting more donors than were actually used. Similar processes will be used for collecting additional samples for the 1000 Genomes Project.

Among the populations whose DNA will be sequenced in the 1000 Genomes Project are: Yoruba in Ibadan, Nigeria; Japanese in Tokyo; Chinese in Beijing; Utah residents with ancestry from northern and western Europe; Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Toscani in Italy; Gujarati Indians in Houston; Chinese in metropolitan Denver; people of Mexican ancestry in Los Angeles; and people of African ancestry in the southwestern United States.

"This project reinforces our commitment to transform genomic information into tools that medical research can use to understand common disease," said Jun Wang, Ph.D., associate director of BGI Shenzhen, whose laboratory will participate in the 1000 Genomes Project and which also took part in the HapMap Project. "It will benefit all nations by creating a valuable resource for researchers around the globe."

The detailed map of human genetic variation will be used by many researchers seeking to relate genetic variation to particular diseases. In turn, such research will lay the groundwork for the personal genomics era of medicine, in which people routinely

will have their genomes sequenced to predict their individual risks of disease and response to drugs.

The data generated by the 1000 Genomes Project will be held by and distributed from the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI), which is part of NIH. There will also be a mirror site for data access at BGI Shenzhen. In addition to a catalog of variants, the data will include information about surrounding variation that can speed identification of the most important variants.

To learn more about the 1000 Genomes Project as the effort develops, and to read *A Workshop to Plan a Deep Catalog of Human Genetic Variation,* which summarizes the meeting that laid the groundwork for the project, go to www.1000genomes.org.

###