Embargoed until 14:30 CEST European time, 13:30 BST UK , 8:30 Eastern US summer time

Contacts:

Louisa Wood or Katrina Pavelin, EMBL–EBI
louisa@ebi.ac.uk
katrina@ebi.ac.uk
+44 (0)1223 494665

Sonia Furtado, EMBL
+49 6221 387 8263
sonia.furtado@embl.de

Don Powell, Wellcome Trust Sanger Institute
+44 (0)12243 496928
Press.officer@sanger.ac.uk

1000 Genomes Project Releases Data from Pilot Projects on Path to Providing Database for 2,500 Human Genomes

Freely available data supporting next generation of human genetic research

The 1000 Genomes Project, an international public–private consortium to build the most detailed map of human genetic variation to date, announces the completion of three pilot projects and the deposition of the final resulting data in freely available public databases for use by the research community. In addition, work has begun on the full-scale effort to build a public database containing information from the genomes of 2,500 people from 27 populations around the world.

Launched in 2008, the 1000 Genomes Project first conducted three pilot studies to test multiple strategies to produce a catalog of genetic variants that are present in 1 percent or greater frequency in the different populations chosen for study (European, African and East Asian). Disease researchers will use the catalog, which is being developed over the next two years, to study the contribution of genetic variation to illness. In addition to distributing the results on the Project's own web sites, the pilot data set is available via the Amazon Web services (AWS) computing cloud to enable anyone to access this unprecedentedly large data set, even if they do not have capacity to download it locally.

"I am indebted to all the project researchers who are making this collaboration so successful," said Richard Durbin, Ph.D., of the Wellcome Trust Sanger Institute, who is co-chair of the consortium. "In the pilot projects we have made significant progress in optimizing the use of next generation sequencing platforms to study human genetic variation, and we can now apply what we have learned to accelerate our efforts to sequence this reference collection of human genomes."

"Completing the goals of the initial pilot projects has been critical to informing how to apply next-generation sequencing in human genetic research, and provides a solid foundation the next

stage of the project," said David Altshuler, M.D., Ph.D., of the Broad Institute, Cambridge, Mass., and co-chair of the project consortium. "We are eager to make rapid progress on the full set of 2,500 genomes and to provide the resulting data for use by the disease genetic community. I fully expect that these data will more precisely define genetic risk factors already discovered, and lead to the discovery of many new risk factors for disease."

A previous public project, the International HapMap Project, provided an initial database of over 3 million human DNA variants present in 270 DNA samples. Information and methods developed by the HapMap Project fueled a first generation of so-called "Genome Wide Association Studies" (abbreviated GWAS) that have localized over 600 novel genetic risk factors for common diseases such as diabetes, heart attack, inflammatory bowel disease, breast cancer, schizophrenia, and other disorders. These studies were limited by technology, however, to studying a subset of more common DNA variants (those with frequency greater than 5-10%).

The 1000 Genomes Project exploits next-generation DNA sequencing technologies to develop a much more complete database – one that goes much lower in frequency, and one that is extended to more human populations. This database will contain all forms of variation – single letter changes (termed SNPs), small insertions and deletions (termed "indels") and large changes in the structure and copy number of chromosomes (termed "copy number variations"). This integrated map is a novel contribution, as previous studies have focused exclusively on one form of DNA variation (even though each of our genomes contains all variety of variation).

"The increased resolution of the 1000 Genomes map will provide researchers with far more detailed sequence information beyond common variants, including millions of less-common and rare variants", said Elaine Mardis, Ph.D., co-director of the Washington University Genome Center and member of the project steering committee. "Researchers who have found regions of the genome associated with disease will be able to look at this data to see an almost complete set of genetic variants in those regions that might contribute directly to disease."

A critical new component of the Project is the selection of 2,500 DNA samples from 27 populations around the world. Each participant has provided explicit consent for full and public release of DNA samples and full sequence data (including recognition of potential risks). The free and public availability of Project data will fuel development of new methods and new approaches to genetic research – applications that would happen much more slowly (if at all) if there were only disease-specific datasets that can't be shared freely on the web (due to more restrictive informed consent).

"We are committed to make these data public to make certain that any institution or researcher around the world can access and work with our datasets to better understand common disease," said Jun Wang, Ph.D., associate director of the Beijing Genomics Institute in Shenzhen, China, and member of the 1000 Genomes Project steering committee. "We must work together if we are going to find those subtle differences in the human genome that lead to diseases like cancer and diabetes."

The uses of Project data will be many. One clear use is to track down the causal mutations underlying initial localizations from GWAS. A second is making it possible to test less common

DNA variants for contributions to disease. And a third is to help identify rare mutations that cause strongly inherited diseases: in studies aiming to find such rare mutations, it is very helpful to have a complete database of common variants that can be screened out to focus attention on those mutations that are unique to an individual or family.

But before such uses could be realized, many technical and analytical challenges had to be overcome. These were the focus of the pilot projects.

Pilot projects – testing essential aspects of project feasibility

The first pilot project involved sequencing the genomes of six people (two nuclear families each with two parents and a daughter) at high coverage. Each sample was sequenced an average of 20 - 60 times, and using a variety of sequencing technologies. Previous "personal genomes" were each based on only a single sequencing method, and thus were limited to what that method could detect. By using multiple methods, the Project has uncovered not only a more complete picture of DNA variation in these individuals, but also learned about the strengths and limitations of each of the current technologies. These data also served as a comparison group for the genome sequences analyzed in the other pilot projects. The six genomes were sequenced by academic centers in China, Germany, the U.K., and the U.S., as well as by three companies, using platforms from the companies: 454 Life Sciences, a Roche company; Applied Biosystems, an Applera Corp. business; and Illumina Inc. All of the platforms were able to sequence 85-90 percent of a genome and produce high-quality data.

The second pilot project sequenced the genomes of 179 people at low coverage -- an average of three passes of the genome. Although sequencing costs are dropping, it is still very expensive to sequence the genomes of hundreds of people deeply enough to find all of the genetic variants in each genome accurately. An alternative approach is to sequence many genomes at light coverage, and then combine the data from many people to discover genetic variants that they share. The results of the pilot project confirmed that this strategy is effective and will allow the project to meet its goal of discovering sequence variants that are shared with other people.

The third pilot project involved sequencing the coding regions, called exons, of 1,000 genes in about 700 people to explore how best to obtain a detailed catalog in the approximately 2 percent of the genome that is composed of protein-coding genes. This Project provided unprecendented sample size to learn about the patterns of rare variation in the human population.

Data analysis and access – and first major release of biomedical data on the Amazon Web Services Cloud

The amount of data produced by the 1000 Genomes Project is unprecedented in biomedical research. Currently, the total size of the datasets is over 50 terabytes, or 50,000 gigabytes. That corresponds to almost eight trillion DNA base pairs, or terabases, of sequence data. Early in the project, merely copying the vast quantities of data between the European Bioinformatics Institute (EBI) in the U.K. and National Center for Biotechnology Information (NCBI), part of the U.S. National Library of Medicine in the U.S. consumed large fractions of both groups' capacity on the Internet for several days.

Researchers can freely access the 1000 Genomes Project pilot data through the 1000 Genomes web site, www.1000genomes.org. Researchers can download the data from NCBI at: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ or from the EBI at: ftp://ftp.1000genomes.ebi.ac.uk/.

For many researchers and institutions, especially those who lack the computer and analytical power to study such a massive data set, an economical option is being tested to access and analyze the pilot data. The pilot datasets of the 1000 Genomes Project (7.3TB of data) are now available as a public dataset through Amazon Web Services (AWS) and integrated into the company's Elastic Compute Cloud (Amazon EC2 and Simple Storage Service, S3) As new data become available and usage of this data increase on AWS, it is anticipated that additional data sets will be available in AWS.

The cost to researchers for computing through Amazon EC2 can be counted in tens of dollars per day compared to the hundreds of thousands of dollars it would cost to purchase the computer infrastructure needed to download and analyze this amount of data locally. Because 1000 Genomes Project data are publicly available from EBI and NCBI, other companies that provide similar computing services are also free to download and provide the data to their clients.

"The 1000 Genomes Project has a simple goal: peer more deeply into the genetic variations of the human genome to understand the genetic contribution to common human diseases," said Eric D. Green, M.D., Ph.D., director of the National Human Genome Research Institute, which provides major funding to the effort. "I am excited about the progress being made on this resource for use by scientists around the world and look forward to seeing what we learn from the next stage of the project."

Consortium researchers are writing a paper that describes the pilot data and the design of the full project that is expected to be published in a peer-reviewed scientific journal later this year.

###

Notes to Editors
About the 1000 Genomes Project

The genetic blueprints, or genomes, of any two humans are more than 99 percent the same. Still, the small fraction of genetic material that varies among people holds valuable clues to individual differences in susceptibility to disease, response to drugs and sensitivity to environmental factors.

The 1000 Genomes Project builds upon the International HapMap Project, which produced a comprehensive catalog of common human genetic variation showing how the variation is organized into chromosome neighborhoods called haplotypes. The HapMap catalog laid the foundation for the recent explosion of genome-wide association studies that have identified more than 650 genetic variants associated with a wide range of common diseases, including type 2 diabetes, coronary artery disease, prostate and breast cancers, rheumatoid arthritis, inflammatory bowel disease and a number of mental illnesses.

The HapMap used genotyping technology and analyzed the common genetic variants in samples from 270 people. The variation it catalogued directly was just a small fraction of the genetic variation within the human genome, although the data provided information about patterns of variation across almost the entire the genome. In contrast, the 1000 Genomes Project catalog will be built by sequencing stretches of the DNA letters, or bases, of 2500 individual genomes. The increased resolution will enable the 1000 Genomes map to provide researchers with far more detailed sequence information than the HapMap, including not only all common variants, but also many less-common and rare variants. Differences in the spelling of the DNA letters as well as larger insertions, deletions, and rearrangements of chunks of the DNA sequence will be captured.

The populations that are expected to included in the Project come from around the world: African-American, Ahom (India), Barbadian, British, Colombian, Dai Chinese, Finnish, Gambian, Han Chinese, Japanese, Kayastha (India), Kinh Vietnamese, Luhya (Kenya), Malawian, Maratha (India), Mexican-American, Peruvian, Puerto Rican, Punjabi (Pakistan), Reddy (India), Southern Han Chinese, Spanish, Tuscan, Utah residents, and Yoruba.

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 have their genomes sequenced to find their genetic variants, which will be the basis for predicting their individual risks of disease and response to drugs.

Organizations that have committed major support to the project are 454 Life Sciences, a Roche company, Branford, Conn.; Life Technologies Corporation, Carlsbad, Calif.; Beijing Genomics Institute, Shenzhen, China; Illumina Inc., San Diego; the Max Planck Institute for Molecular Genetics, Berlin, Germany; the Wellcome Trust Sanger Institute, Hinxton, Cambridge, U.K.; and the NHGRI, which supports the work being done by Baylor College of Medicine, Houston, Texas; the Broad Institute, Cambridge, Mass.; and Washington University, St. Louis, Missouri. Researchers at many other institutions are also participating in the project including ones in Barbados, China, Colombia, Finland, the Gambia, India, Malawi, Pakistan, Peru, Puerto Rico, Spain, the U.K., the U.S., and Vietnam.

Additional information about the project, including a list of all participants and organizations, can be found at http://www.1000genomes.org/.

The Wellcome Trust is a global charity dedicated to achieving extraordinary improvements in human and animal health. It supports the brightest minds in biomedical research and the medical humanities. The Trust's breadth of support includes public engagement, education and the application of research to improve health. It is independent of both political and commercial interests. www.wellcome.ac.uk

The Wellcome Trust Sanger Institute, which receives the majority of its funding from the Wellcome Trust, was founded in 1992 as the focus for UK sequencing efforts. The Institute was responsible for the completion of the sequence of approximately one-third of the human genome as well as genomes of model organisms such as mouse and zebrafish, and more than 90 pathogen

genomes. In October 2005, funding was awarded by the Wellcome Trust to enable the Institute to build on its world-class scientific achievements and exploit the wealth of genome data now available to answer important questions about health and disease. These programmes are built around a Faculty of more than 40 senior researchers. The Wellcome Trust Sanger Institute is based in Hinxton, Cambridge, UK.
http://www.sanger.ac.uk/

EMBL-European Bioinformatics Institute:
The European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL) and is located on the Wellcome Trust Genome Campus in Hinxton near Cambridge (UK). The EBI grew out of EMBL's pioneering work in providing public biological databases to the research community. It hosts some of the world's most important collections of biological data, including DNA sequences (ENA), protein sequences (UniProt), animal genomes (Ensembl), three-dimensional structures (the Protein Databank in Europe), data from gene expression experiments (ArrayExpress), protein-protein interactions (IntAct) and pathway information (Reactome). The EBI hosts several research groups and its scientists continually develop new tools for the biocomputing community.

European Molecular Biology Laboratory
The European Molecular Biology Laboratory is a basic research institute funded by public research monies from 20 member states (Austria, Belgium, Croatia, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom) and associate member state Australia. Research at EMBL is conducted by approximately 85 independent groups covering the spectrum of molecular biology. The Laboratory has five units: the main Laboratory in Heidelberg, and Outstations in Hinxton (the European Bioinformatics Institute), Grenoble, Hamburg, and Monterotondo near Rome. The cornerstones of EMBL's mission are: to perform basic research in molecular biology; to train scientists, students and visitors at all levels; to offer vital services to scientists in the member states; to develop new instruments and methods in the life sciences and to actively engage in technology transfer activities. Around 190 students are enrolled in EMBL's International PhD programme. Additionally, the Laboratory offers a platform for dialogue with the general public through various science communication activities such as lecture series, visitor programmes and the dissemination of scientific achievements.

NHGRI is one of 27 institutes and centers at the NIH, an agency of the Department of Health and Human Services. The NHGRI Division of Extramural Research supports grants for research and for training and career development at sites nationwide. Additional information about NHGRI can be found at its Web site, www.genome.gov. The National Institutes of Health — "The Nation's Medical Research Agency" — includes 27 institutes and centers, and is a component of the U.S. Department of Health and Human Services. It is the primary U.S. federal agency for conducting and supporting basic, clinical and translational medical research, and it investigates the causes, treatments and cures for both common and rare diseases. For more, visit www.nih.gov.