

1000 Genomes Project

 consortiapedia.fastercures.org/consortia/1gp/

Research Areas



Tool Development



Biomarker Research

Diagnostic, Genomic Biomarker



Data-Sharing Enabler

At a Glance

- Status: **Active Consortium**
- Year Launched: **2008**
- Initiating Organization: **National Institutes of Health/National Human Genome Research Institute**
- Initiator Type: **Government**
- Location: **International**

Abstract

The 1000 Genomes Project is a consortium focused on developing methods to collect, share, and integrate genomic data generated from multiple sources in multiple countries, in an effort to provide a foundation for investigating the relationship between genotype and phenotype. The goal is to use these methods to create a catalog of common genetic variation across human populations, which will be made accessible for the broader scientific community.

Mission

The aims of the 1000 Genomes Project are to genotype and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations. To achieve these aims, the consortium worked across its partners to develop and evaluate methods for high-throughput sequencing. This process included the development of robust protocols for generating whole-genome shotgun and targeted sequence data, as well as algorithms to detect variants. In addition to a data repository, the study results should serve as a template for future genome-wide sequencing studies on larger sample sets.

The consortium identified the following populations for DNA sequencing: 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.

Consortium History

The 1000 Genomes Project was launched in 2008 and conducted three pilot studies to test multiple strategies for producing a catalog of genetic variants with a frequency of at least 1% within the different study populations (European, African, and East Asian). The data of approximately 2,500 human genomes were made available to the public in 2010 via the consortium's two websites and Amazon Web Services (AWS). The consortium leverages the findings from the International HapMap Project, a comprehensive catalog of common human genetic variation from 270 people that shows how genomic variation is organized into chromosome neighborhoods called haplotypes.

The first pilot project sequenced 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, using multiple technologies. By using multiple technologies, the project uncovered a more complete picture of DNA variation in these individuals and identified the strengths and limitations of each technology. The data also served as a comparison group for the genome sequences analyzed in the other pilot projects. The second pilot project sequenced the genomes of 179 people at low coverage—an average of three passes of the genome. This project's results were used to confirm the efficacy of this methodology. The third pilot project sequenced the coding regions, called exons, of 1,000 genes in about 700 people to explore how best to obtain a detailed catalog for the approximately 2% of the genome that consists of protein-coding genes.

Structure & Governance

Decisions in all activities require consensus by the 1000 Genomes Steering Committee and the 1000 Genomes Analysis Group. The Steering Committee serves as the main governing board of the 1000 Genomes Consortium and includes the Project co-chairs, working group co-chairs, a representative

from each sequencing center, and some additional members, including a program director from the National Institutes of Health's (NIH) National Human Genome Research Institute (NHGRI). Almost all calls by the Steering Committee are open to all Project participants.

The Analysis Group includes all the consortium participants working on data processing and analyses, as well as the sponsors and funders (Scientific Management team). The Principal Investigator (PI) and relevant staff from an awarded cooperative agreement may be added to the Analysis Group, depending on the needs of the Project; awardees not part of the Analysis Group are still expected to work closely with the group. Awardees are required to accept and implement the common guidelines, procedures, and policies approved by the Steering Committee and the Analysis Group. Each awardee PI has one vote on the Analysis Group.

Scientific Management team members are scientists who provide normal stewardship of the funding and provide scientific and programmatic involvement through technical assistance, advice, and coordination. They also participate as members of the Steering Committee and Analysis Group.

The Steering Committee and the Analysis Group have the authority to add additional members. The Steering Committee and the Analysis Group can also establish ad hoc groups, which would include representatives from the grantees and the funding agencies, and possibly other experts. For NIH-funded awardees, any disagreements that arise in scientific or programmatic matters between award recipients and the NIH may be brought to arbitration. If this occurs, then an Arbitration Panel will be convened with three members: one designee of the Steering Committee chosen without NIH staff voting, one NIH designee, and one designee with expertise in the relevant area who is chosen by the other two; in the case of individual disagreement, the first member may be chosen by the individual awardee.

Sequencing work was carried out at the Sanger Institute, BGI Shenzhen, and NHGRI's Large-Scale Sequencing Network, the latter of which includes the Broad Institute, 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. All samples from the 1,000 genomes are available as lymphoblastoid cell lines (LCLs) and LCL-derived DNA from the Coriell Cell Repository as part of the NHGRI Catalog. In addition Standard Population DNA Panels for the 1000 Genomes and HapMap projects are available at \$1,000 or less each.

Financing

With an estimated cost of \$30-50 million, the Project received the majority of its support from the Wellcome Trust Sanger Institute, Beijing Genomics Institute, and NHGRI. Three industrial sequencing companies—454 Life Sciences (Roche), Life Technologies, and Illumina—provided in-kind sequence data and capacity, estimated to be worth approximately \$700,000 for the pilot phase, with more sequencing contributions anticipated for the full project.

Intellectual Property

There is no associated intellectual property—all data are freely available to the public via the Project websites and the cloud via AWS.

Data Sharing

The Project samples are mostly anonymous and have no associated medical or phenotype data; for some of the populations, the collectors have phenotype data but these data are not at Coriell Institute and are not distributed. All samples from the 1,000 genomes are available as LCLs and LCL-derived DNA from the Coriell Cell Repository as part of the NHGRI Catalog.

NIH's National Center for Biotechnology Information (NCBI) and European Bioinformatics Institute began making Project data freely available to researchers in 2008. In addition to information about catalog variants, the data will include information about surrounding variation that can speed identification of the most important variants. In 2010, the pilot data became available via the AWS cloud for free, enabling any researcher to access and analyze the data at a fraction of the cost it would take for his or her institution to acquire the needed internet bandwidth, data storage, and analytical computing capacity. The data can be seamlessly accessed through services such as Amazon Elastic Compute Cloud and Amazon Elastic MapReduce, which provide organizations with the highly scalable resources needed to power computing applications often needed in research. Researchers pay only for the additional AWS resources needed to further process or analyze the data. In 2012, additional data were released via AWS, including results from sequencing the DNA of some 1,700 people. Data are available via the consortium website at <http://www.1000genomes.org>, as well as the NCBI website at <ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes> and the Wellcome Trust website at

<ftp://ftp.1000genomes.ebi.ac.uk>.

The data producers aim to release the Project data quickly, prior to publication, with the expectation that they will be valuable for many researchers. In keeping with Fort Lauderdale principles, data users may use the data for many studies, but they are expected to allow the data producers to make the first presentations and to publish the first paper with global analyses of the data.

Impact/Accomplishment

This effort has already resulted in the completion of three pilot projects, and the resulting data have been deposited 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. This unprecedented large data set is available on the Project's own web sites, but it is also available via the AWS computing cloud for researchers who may not have capacity to download it locally.

In addition to sequencing the data, the pilot projects developed criteria to determine the strengths and weaknesses of current sequencing technologies, demonstrated a method for sequencing at low coverage, and found methods to perform high-throughput exon sequencing for large populations.

Links/Social Media Feed

Homepage	http://www.1000genomes.org/
Twitter	@1000genomes
Other website	http://aws.amazon.com/1000genomes/
Other website	ftp://ftp.1000genomes.ebi.ac.uk
Other website	ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes

Points of Contact

Central email: info@1000genomes.org

Project Co-Chairs:

Richard Durbin, PhD (rd@sanger.ac.uk)

Wellcome Trust Sanger Institute

Wellcome Trust Genome Campus

David Altshuler, MD, PhD (altshuler@molbio.mgh.harvard.edu)

Associate Professor of Genetics and of Medicine

Harvard Medical School

Massachusetts General Hospital

NHGRI Staff:

Lisa D. Brooks, PhD (lisa.brooks@nih.gov)

Program Director

Genetic Variation Program

National Human Genome Research Institute

National Institutes of Health

Adam Felsenfeld, PhD (adam.felsenfeld@nih.gov)

Program Director

Large-Scale Sequencing Program

National Human Genome Research Institute

National Institutes of Health

Jean McEwen, JD, PhD (mcewenj@mail.nih.gov)

Program Director

Ethical, Legal, and Social Implications Program

National Human Genome Research Institute

National Institutes of Health

Sponsors & Partners

Affymetrix, Inc.

Albert Einstein College of Medicine

Baylor College of Medicine

BGI-Shenzhen

Bilkent University, Turkey
Boston College
Brigham and Women's Hospital
Broad Institute
Cardiff University
Cayetano Heredia University
Chinese Academy of Sciences
Complete Genomics, Inc.
Coriell Institute
European Bioinformatics Institute
European Molecular Biology Laboratory
F. Hoffmann-La Roche Ltd. / 454 Life Sciences
Federal Ministry of Education and Research (BMBF,
Germany)
Genetic Alliance UK
Gregor Mendel Institute
Illumina, Inc.
Imperial College London
Johns Hopkins University
Leiden University
Life Technologies Corp.
Louisiana State University
Massachusetts General Hospital
Max Planck Institute for Molecular Genetics
McGill University
Medical Research Council (UK)
Mount Sinai School of Medicine
National Institutes of Health/National Center for
Biotechnology Information
National Institutes of Health/National Human Genome
Research Institute
National Institutes of Health/National Institute of
Environmental Health Sciences
National Planning and Development Committee (China)
Oxford University
Ponce School of Medicine and Health Sciences

Rutgers University
Shenzhen Local Municipal Government (China)
Simon Fraser University
Stanford University
Translational Genomics Research Institute
Tulane University
University College of London
University of Arizona
University of Barcelona
University of California, Los Angeles
University of California, San Diego
University of California, San Francisco
University of California, Santa Cruz
University of Chicago
University of Copenhagen
University of Geneva
University of Maryland
University of Medicine and Dentistry of New Jersey
University of Michigan
University of North Carolina, Chapel Hill
University of Puerto Rico
University of Texas
University of Texas/MD Anderson Cancer Center
University of the West Indies
University of Utah
University of Virginia
University of Washington
Virginia Tech
Washington University in St. Louis
Wellcome Trust
Wellcome Trust Centre for Human Genetics
Wellcome Trust Sanger Institute
Yale University

Updated: 06/20/2017