Plan for Funding
Analysis to Support Data Production and
Analysis of Data from the 1000 Genomes Project

National Advisory Council for Human Genome Research, September 2008

**Purpose:** At the May 2008 meeting of the National Advisory Council for Human Genome Research, Council requested that NHGRI staff develop a plan for providing funding for the analysis component of the 1000 Genomes Project. The Council recommended that NHGRI should provide such funding in a way that balances the analytical activities necessary to address the project's short-term design issues and production needs with an opportunity for competition for funds for longer-term analysis needs. The following proposal comprises a three-pronged approach to funding the analysis needs of the 1000 Genomes Project.

**Background:** The 1000 Genomes Project aims to support genome-wide association (GWA) studies by providing a public resource of almost all genetic variants across the human genome with a frequency of 1% or higher, and of genetic variants with even lower frequencies in gene regions. The initial strategy for doing this proposes to sequence 1200-1500 samples to a depth of coverage of at least 4X across the genome and 12X in gene regions. The project has begun with three pilot projects to test this strategy and to address several issues related to the design of the full-scale project. The data from these pilots will be analyzed and the design for the full-scale project will be developed at a meeting in mid-November of this year.

The 1000 Genomes Project is the first large-scale project that uses the next-generation sequencing platforms for extensive human resequencing. The Project has developed rapidly since its adoption at an initial international meeting in September of 2007. The large-scale sequencing centers, both here and abroad, were able to quickly re-allocate sequence production to this project. They started producing sequence data for the pilot projects in January of this year and plan to finish this fall. The full project sequencing should start this fall and continue for about two years.

The pilot projects have critical needs for data analysis to support the production and interpretation of the pilot dataset; these analyses require funding. The next-generation sequencing methods are still in an early stage of development for production use, so many analyses still need to be developed to produce the data. Some of these, such as developing quality scores, estimating error rates, and providing processed datasets including genotypes for individual samples, will apply to many types of sequencing projects. Other analyses will be more specific to the 1000 Genomes Project, such as using the pilot data to develop the design for the full project. Pipelines for quality assessment and data integration will need to be established and monitored to produce the full project dataset. Finally, this dataset will require many analyses to characterize it, such as for distributions of coverage and allele frequency.

Analysis work has already begun in the 1000 Genomes Project pilots. Some of the groups involved have funding from NIH, the Sanger Institute, the Wellcome Trust, or China to work on specific analysis issues. Other critical analysis work, however, is as yet unfunded. All data analysis activities for the 1000 Genomes Project are being tightly coordinated through the project Production and Analysis Groups.

**Plan:** Staff proposes three ways for NHGRI to fund the additional analyses needed to accomplish the goals of the 1000 Genomes Project. Together, these proposals attempt to balance the immediate needs of the project with an opportunity to support an open competition for new participants and ideas:

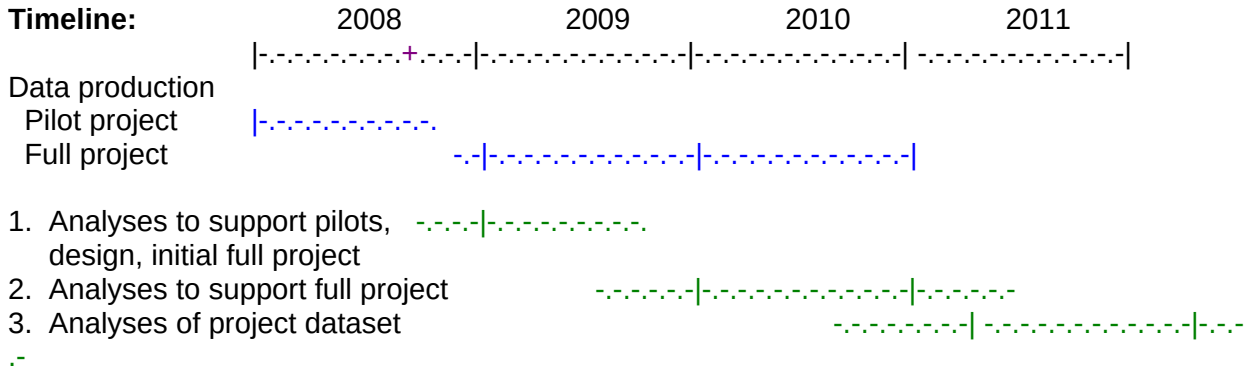
1. ***Analyses to support the production of pilot data, design of the full project, and early full project data production*.** Some of the groups that are actively working on the pilot project analyses have funding for these activities (Appendix A). However, others require funding. As noted in the Background section, all of these analyses are being done in close coordination with the Project's Production and Analysis Groups and with the Steering Committee. See Appendix B (to be provided) for the additional analyses that need to be done. Staff proposes to provide funding for the groups that need it for these analyses through supplements (to be discussed in Closed Session). This funding would start in the fall of 2008 and last for one year.

2. ***Analyses to support data production for the full project.*** Data for the full project should start to become available late in 2008, and analyses to support the production of those data will need to be funded. Staff proposes a limited competition RFA, which would be sent to all of the U.S. participants in the current 1000 Genomes Analysis Group, to provide funding for the duration of the full project based on competition and peer review. This funding would start about July 2009 and last until the completion of the production of the sequence data and processed data in early 2011. The details are presented in Concept Clearance (I).

Rationale for use of a limited competition approach: By July of 2009, which is the earliest that these awards could be made, several groups will already have gained a considerable amount of experience in understanding the data in minute detail and in developing the data-processing pipelines. They will, therefore, be in the best position to monitor the data and to develop new analyses that the initial work shows is needed. It would be inefficient to switch out those groups for new groups, which would have to start from scratch and develop their own pipelines for the same analyses. However, it will be important to have some groups within the project, who are already familiar with the data, compete to do the needed activities in the context of the plan for the full project that is to be developed at the project meeting in mid-November 2008.

3. ***Analyses of the dataset produced by the full project.*** Beyond the analyses needed to produce the 1000 Genomes data, an RFA open to all investigators would solicit applications broadly for analysis of the complete project dataset. This funding would start in early 2010 and last for two years. This is the subject of Concept Clearance (II).

**Timeline:**



## Appendix A

### Funded Analysis Efforts for Production and Evaluation of the Pilot Data and Design of the Full 1000 Genomes Project

Some of the infrastructure and analyses to support data production for the 1000 Genomes Project are already in place or are being funded. These include:

A. The Data Coordination Center (DCC). The 1000 Genomes DCC is provided jointly by the EBI and NCBI, with funding from the Wellcome Trust and NIH. The DCC accepts the trace data from the sequencing centers, exchanges data between the two DCC sites, tracks the data, and provides a web site. When the Analysis Group has developed standard procedures, the DCC will implement central processes for filtering and mapping the reads, assessing data quality, calling variants, inferring haplotypes, calculating LD, and providing other data types as needed.

B. The Short-Read Archive (SRA) and the tools to transfer data from the machines to the SRA. The NCBI developed the SRA, and EBI is developing a similar SRA. The informatics staffs at the sequencing centers have developed automated tools to transfer the data from the sequencing instruments to the SRA database.

C. Mapping the reads. The sequencing centers and the platform companies have developed methods to map the reads onto the reference human genome. Richard Durbin of the Sanger Institute has mapped the reads and given feedback to the sequencing centers on each submission. His group is comparing several mapping programs (MAQ (Durbin), MOSAIK (Marth), ELAND (Illumina), and SOAP (Wang)) to determine which one, or combination, should be centrally implemented at the DCC. The Sanger group is also considering the quality of the read mapping.

D. Base-calling. The companies and sequencing centers have developed methods to call bases from the short-read data and assess their accuracy. Gabor Marth is developing better methods with R01 funding from NHGRI. He is working with all the companies to improve the calling of SNPs and indels.

E. Structural variants. The groups of Evan Eichler, Jonathan Sebat, and Charles Lee have R01 funding from NHGRI to examine structural variation. Matthew Hurles has funding from the Sanger Institute for this work. These groups are coordinating their work as part of the 1000 Genomes Project. The groups are using the datasets they produced on fosmid paired-end libraries and with dense arrays to validate the 1000 Genomes data from the pilot projects, to calibrate error rates, to provide input on the design of the full project as it relates to structural variants, and to monitor the structural variation data for the full project. This will involve some difficult *de novo* sequence assembly in regions with complex structural variants.

F. Simulations of the data. Carlos Bustamante has R01 funding from NHGRI to analyze sequence data; his group has been developing simulated datasets for evaluating the pipelines for processing the sequence data, for interpreting the observed data, and for helping to decide on the design of the full project from the pilot data.

G. Gene-region data. Andrew Clark has R01 funding from NHGRI to examine the gene regions that will be sequenced deeply as part of the pilot projects, to address how the platforms work and the design of the full project.

## CONCEPT CLEARANCE (I)

## LIMITED COMPETITION RFA TO FUND THE ANALYSIS COMPONENT OF THE FULL-SCALE 1000 GENOMES PRODUCTION EFFORT

According to the current timetable, data for the full-scale 1000 Genomes Project will start being produced by the end of 2008 and will continue until about the end of 2010. Once the initial pipelines for data production and quality assessment are put in place, many analyses will still be needed to support the production of the public dataset from the full-scale project. The groups that have been involved during the pilot phase in developing these pipelines will be the most familiar with their details and, therefore, will be the best able to scale up these analyses efficiently, along with the scaled data production effort. Accordingly, a limited competition (letter RFA) is proposed, to allow those U.S. investigators who have been working on the project as part of the Analysis Group to apply for funding for the continuing work. About 15 PIs from 10 institutions would be eligible to apply. These applications will be peer-reviewed. Analyses that will be needed to support data production include:

A. Monitoring data quality. Each sequencing center will monitor its own data quality. However, the project Production Group working with the Analysis Group will develop the standard methods that each center will use. The data will need to be monitored on an ongoing basis to ensure that the different platforms and centers are producing high-quality data both individually and when integrated. Long reads and paired-end data from some platforms will make up for some of the limitations of the short-read data, and the coverage of the various types of sequence data will need to be continually monitored across the genome, particularly for structurally complex regions. The monitoring will ensure that the data are of at least as high quality as the initial data and models, and continue to be adequate to meet the goals of the project.

B. Validating variants. Samples of variants of different frequencies and types will need to be validated by genotyping or additional sequencing to provide experimental estimates of false positive and negative rates. The numbers for the various categories will need to be determined, and the validation experiments will need to be interpreted.

C. Combining data across samples. The central analysis method of the project will be to combine data across samples to find variants that are present down to low frequencies. This sharing of data across samples means that although individual samples will not be genotyped with high confidence, the variants will be found with high confidence using the entire dataset. Methods for combining data across samples will need to be fully developed, applied to the large amounts of data that will be produced, and evaluated for effectiveness across the genome and in gene regions. The depth of coverage that provides sufficiently accurate variant discovery will need to be monitored to know when the project has reached its goals.

D. Producing genotypes, haplotypes and other processed data types for release to the community. The methods for base-calling will need to be fully implemented, and no doubt improved as more data become available, as will methods used to infer genotypes, haplotypes, and LD patterns. As more data are generated the structural variants will have to be integrated with SNP variants.

E. Developing tools to work with the data. The massive amounts of data that will be produced will require new tools to transfer, visualize, simulate, and summarize the data.

F. Developing additional analyses. It is anticipated that additional needs will be identified during the course of the full-scale production phase and that new analysis methods will need to be developed in a timely manner.

The limited competition RFA would be released in September of 2008. Eligible applicants would be investigators at U.S. institutions who are in the 1000 Genomes Analysis Group and do not already have sufficient funding to support their participation in the analysis efforts for the full-scale project. Applications would be received in December 2008, reviewed in the Spring of 2009, discussed at May 2009 Council, and funded in July 2009 for two years. Data production for the full-scale phase of the 1000 Genomes Project is anticipated to be complete by about the end of 2010. The timing for this analysis work would allow awardees to complete the necessary analyses shortly after the completion of the data production phase. U01s are proposed, because the analyses will need to be closely coordinated with the project.

The total amount of funding would be up to $3 million for each of two years. Direct costs for each year in an application should be no more than $500,000. It is anticipated that 4-6 awards will be funded. Awards will be made only to U.S. institutions.

This RFA would be restricted to support of efforts to produce the 1000 Genomes data and make them available to the research community in a useful form. It would not provide funding for analyses of particular genomic regions of specific disease interest, follow-up analysis of GWAS datasets, additional sequencing or genotyping, other experimental data validation, or sample collection.

## CONCEPT CLEARANCE (II)

## RFA TO SOLICIT PROJECTS TO ANALYZE THE COMPLETE 1000 GENOMES DATASET

When the project has produced the full 1000 Genomes dataset, many analyses will be needed to assess its quality, characterize it, provide additional processed datasets, do a number of global analyses, and develop better tools for using the dataset to study biomedical and biological research problems. An RFA is proposed for an open competition to do these analyses.

Examples of the types of analyses that will be needed include, but are not limited to:

A. Analysis of genomic coverage. The depth of sequence coverage across the genome and in gene regions will affect the data quality. Random and systematic biases will produce different types of errors.

B. Analysis of data quality. This will include the accuracy of base calls as a function of factors such as coverage, variant allele frequency, and genomic context. Various types of structural variants will differ in their rates of false positives and negatives.

C. Analysis of the frequency distribution of variants. A major goal of the 1000 Genomes Project is to find rare variants, so it will be important to examine their frequency distributions genome-wide and in gene regions.

D. Analysis of haplotypes and the value of shared data across samples. Sharing haplotype data across samples is a major data analysis strategy for the project for finding variants and inferring haplotypes, so it will be important to assess how well this sharing works.

E. Imputation of variants. Another major goal of the project is to impute variants so that GWA studies can use those data for association analyses. Thus it will be important to examine how well variants of different types and frequencies can be imputed genome-wide and in gene regions.

F. Analysis of special regions. Regions such as HLA, which have extensive sequence diversity and structural variation, will require particular attention.

G. Comparisons of populations. The large population groups to be studied in the project are expected to have some differences in the amount of variation and LD patterns. The value of the populations in providing additional information on rare alleles should be evaluated. The effects of admixture should also be examined.

H. Looking for signals of natural selection. In this large dataset, with many individuals and several populations, many types of signals of natural selection should be detectable.

I. Considering how the resource should be improved. By the time that the dataset is produced, several studies of sequencing for follow-up to GWA studies should have been analyzed. It will be useful to assess how well the data from the 1000 Genomes Project meet the needs of this follow-up for finding almost all variants in regions of GWAS hits, and what additional data will be needed.

J. Tools for accessible analyses of the data for GWA and other studies. The project will produce a massive dataset. Maximizing its value to the scientific community will require the development of many tools to filter, visualize, and examine the data.

The RFA would be released in January of 2009, and applications would be received in June 2009, reviewed in the fall, discussed at February 2010 Council, and funded about April of 2010. The project is anticipated to finish producing the sequence data by about the end of 2010. This timing would allow awardees to become familiar with the data, develop their methods on the initial data from the project, and then be in position to do the full-scale analyses when the full dataset is produced. Two years of funding are proposed, which should be adequate to develop the methods and apply them to the data; the goal is to produce the results for the community quickly. U01s are proposed, because the analyses will need to be closely coordinated with the project. Awardees might be integrated into the Analysis Group of the 1000 Genomes Project, depending on the needs of the project.

The total amount of funding would be up to $4 million for each of two years. Direct costs for each year in an application should be no more than $300,000. It is anticipated that 8-10 awards will be funded. Awards will be made only to U.S. institutions.

This RFA would not provide funding for analyses of particular genomic regions of specific disease interest, follow-up analysis of GWAS datasets, additional sequencing or genotyping, other experimental data validation, or sample collection.