Concept Clearance
NHGRI Large-Scale Sequencing Program May 2010

DRAFT 05032010

This document reflects many discussions beginning with the "Future of the Large Scale Sequencing Program" workshop report: http://www.genome.gov/Pages/Research/DER/DERReportsPublications/SeqPlanWkshopReportFinal.pdf; discussions at two Council meetings (late 2009 and February 2010); a number of discussions with the Scientific Advisors to the Large Scale Sequencing Program (SAP) and subsequent staff planning discussions.

We propose that the NHGRI large-scale sequencing program renewal should consist of the four separate initiatives described below, which together are intended to: take the best advantage of the current scientific opportunities made possible by very high throughput efforts; further develop and leverage sequence analysis tools for the benefit of the wider genomic sequencing community; and stimulate the use of genomic sequencing in patient care.

1. **Large-scale Genome Sequence and Analysis centers** (Cooperative agreements, four years, 80 – 85% of the program). This component would continue the current large-scale centers aspect of the NHGRI sequencing program, with emphasis on high priority, large-scale projects towards a range of biological and biomedical aims. There are a number of major opportunities that can be addressed in the next four years by this component, including understanding the genomic basis of complex disease, for example by sequencing whole genomes of thousands of phenotyped individuals from disease cohorts or prospective studies, and expanded elucidation of the genomic basis for cancer. Projects of this type require a degree of scale and organization that are appropriate for large-scale centers. A summary of additional potential opportunities is appended.

   Together with providing capacity for large, high-priority projects, this component is meant to fund centers that will:
   - continue to set the state-of-the-art in high-throughput genomic sequencing, analysis and project design;
   - be flexible enough with regard to sequencing technologies and analytical capabilities to pursue many different types of project as the scientific opportunities change;
   - pursue projects that use and integrate different genomic data types (genomic, expression, etc.) to best address the scientific opportunities;
   - implement new technologies;
   - maintain sequence finishing capabilities;
   - maintain integrated informatics and analysis pipelines;

- disseminate genomics "know-how" to the community, for example by helping to establish standards for the wider genome sequencing community, developing analysis methods and software, testing project designs, etc.;
- serve as points of organization for the communities that will use the data;

NHGRI will continue to use a range of approaches to identify new projects that are appropriate for this component. Because the science and the sequencing technology are together changing so rapidly, NHGRI expects that there will need to be multiple routes for new projects to be pursued, including:

- A mechanism for communities to propose projects, with approval by committee. This will need to be modified from the current similar mechanism (the Coordinating Committee) to be more transparent with more formal review. It is critical that such projects match the scale that is appropriate for the large centers.
- Collaborations with other NIH institutes, including co-funding in some cases; this can build on the successful collaborations already in progress with NCI, NIMH, and NIDDK.
- NHGRI-initiated projects, identified by program staff through a variety of mechanisms including workshops.
- Center-initiated projects, approved by the Sequencing Advisory Panel (SAP).

2. **A Center for Mendelian Disorders** (Cooperative agreements, four years, 9-10% of the program). Since shortly after the completion of the human genome sequence, we have seen advantages in funding centers in a way that did not constrain their focus to a particular project. However, elucidating the basis for Mendelian disease is now an opportunity that fits well with the concept of a dedicated, mid-sized center. OMIM (as of 4/26/2010) records 1780 Mendelian disorders and 1994 more that are suspected to be Mendelian, with no known molecular basis. Establishing the molecular basis of all of these would be extremely valuable, not only for understanding the basis for each individual disorder, but as a collection of variants with strong phenotypes that will be powerfully informative for human biology. It has recently become feasible to find the molecular basis of Mendelian disorders by sequencing whole exomes of just a few individuals (typically no more than 5). While such studies are individually not expensive in the hands of an expert, the ability to routinely sequence and analyze whole exomes is not yet widespread. Moreover, even if it were, there is a considerable challenge in organizing such efforts on an individual basis --- identifying appropriate samples, now widely dispersed; ensuring that samples are associated with adequate informed consent for genomic sequencing; ensuring that as many of these disorders as possible are worked on without redundancy, and so on. We therefore propose to establish a center whose mission will be to take on the sequencing of Mendelian diseases. Such a center would have to be able to coordinate the availability of samples, maintain a focused, state-of-the-art exome sequencing and analysis capability, be a resource to other investigators interested in pursuing the identification of individual Mendelian disease genes by sequencing, and involve interested disease research communities in the analyses. Such a center, operating at full scale, would be expected to (1) be able to

resolve the molecular basis of on the order of 40-50 disorders per year by itself, and over the term of the grant organize and (2) coordinate the samples for all remaining unsolved Mendelian disorders, by identifying samples within the community, obtaining commitments from the investigators who have samples for distribution of those samples to other groups able to do whole-exome sequencing , establishing lists of samples for community coordination, and for generally facilitating the establishment of the molecular basis for as many of these disorders as possible. This component of the NHGRI sequencing program is likely to attract interest from other NIH institutes with specific disease interests.

3. **Clinical Sequencing Exploration Projects** (R01s, three years, 4-5% of the program). An important outcome of NHGRI's medical sequencing activities will be the application of the knowledge gained to the care of individual patients. Until recently, this still seemed remote — the technologies were too expensive and inefficient to apply to individual patients, and the knowledge base was too poor (we make a loose distinction here between clinical application of individual genetic test results for individual known alleles, which of course has been occurring for some time, and the idea of a more genomic approach). However, both of those circumstances are changing. With this component of the sequencing program, we aim to stimulate the use of sequence information in the clinic to begin to bridge the gap between discovery and personalized medicine. We will solicit approximately 5 projects in which a clinician seeks to obtain substantial sequence and other genomic information (probably from a local core or other sequencing service provider) from patients in order to guide patient care by informing diagnosis, prognosis, or treatment. Collaborative projects with both clinical and genetic expertise may be expected or even necessary. NHGRI will use this opportunity to gain insight into the specific issues that need attention if individualized care is to become useful, including understanding what types of clinical uses represent the best opportunities for application of patient genomic sequencing; project design; and analysis and software capabilities and needs. This initiative also raises significant ethical, legal and social issues, from patient consent to data release to physician/patient communication to regulatory issues. We thus expect that these awards will have a substantial and well-integrated ELSI component.

4. **Robust Genomic Sequence Analysis Tools** (R01's, four years, 4-5% of the program, plus SBIR grants). A strong, consistent recommendation from the large-scale sequencing workshop and from subsequent informatics workshops was that NHGRI should encourage the creation of robust, well-documented and well-supported informatics tools for sequence analysis. While many of these tools have been developed (for example tools for sequence alignment, whole genome assembly, and variant calling), most of them exist only as custom scientific software at the large-scale sequencing centers,. where they work well. However, the centers have not been incented, or are necessarily the proper sites, to make those tools robust, to compare them to establish the "best" tool, to provide adequate documentation or user support, or to engineer them into robust high-performing software tools that can be used by the many groups that are becoming involved in genomic sequencing. This initiative will seek to fund groups that will take scientific sequence analysis software from both the large-scale

centers and from other sources, and make them robust and available broadly. It is not clear whether the best model for this is an academic one or a private (for-profit) one; NHGRI has received conflicting advice on this point. This initiative would thus include both SBIR and normal grant mechanisms.

Finally, there is a question of overall funding for the NHGRI large-scale sequencing program. Currently, the program is funded at ~$110M/year. The initiatives proposed here could benefit from a similar amount, in which case (a) the distribution as proposed would be $90M for initiative 1, $10M for initiative 2, and $5M each for initiatives 3 and 4; and (b) in this formulation, the large-scale sequencing centers would receive a significant funding decrement of about 18% overall in the first year (it should be noted that NHGRI has cut the large-scale center budget by roughly 10% per year since the completion of the human genome sequence). Depending on the priorities that are developed on the basis of the discussion at the May 2010 Council meeting and during the remainder of the NHGRI planning process, both the overall amount of funds made available for large-scale sequencing and the proportion of funds that go to each of the approved initiatives could be adjusted up or down.