National Advisory Council for Human Genome Research

May 16, 2011

Concept Clearances for RFAs

Expanding the Encyclopedia of DNA Elements (ENCODE) in the Human and Model Organisms

A Data Analysis and Coordination Center for the Encyclopedia of DNA Elements (ENCODE) Project

Computational Analysis of the Encyclopedia of DNA Elements (ENCODE) Data

This Concept Clearance is for three solicitations to implement the next phase of the ENCODE Projects:

1. The NHGRI proposes to expand the ENCODE Projects with a goal of moving towards as complete a catalog as is feasible within the limitations of current technology.
2. To facilitate the project, the NHGRI proposes that the storage, display and release of data generated by the ENCODE Projects should be coordinated through a centralized data coordination center. This center will also coordinate the data analysis needs for the ENCODE Project.
3. Lastly, NHGRI proposes to seek applications to support individual research grants to expand the pool of researchers analyzing the ENCODE data.

Background:

The long-term goal of the Encyclopedia of DNA Elements (ENCODE) and the related modENCODE Projects is to generate comprehensive catalogs of functional elements in the human, and *C. elegans* and *D. melanogaster* genomes, respectively. The ENCODE Project was launched in 2003 as a pilot to analyze 1% (30Mb) of the human genome sequence, and was expanded to study the entire human genome in 2007. Also in 2007, comparable genome-wide analyses of the functional elements in the *C. elegans* and *D. melanogaster* genomes were initiated as the modENCODE Project. In 2009, with funds from the American Recovery and Reinvestment Act (ARRA), a small, two-year effort to generate related data from the mouse genome was established. For the purposes of this solicitation, and unless otherwise noted, these efforts will collectively be referred to as the "ENCODE Projects". In parallel with these data production activities, NHGRI has supported several technology development

efforts to expand the repertoire of high-throughput methods available for identifying functional elements, and a new solicitation to continue this effort was approved at the February 2011 meeting of the National Advisory Council for Human Genome Research.

In May 2010, the NHGRI reported to the National Advisory Council for Human Genome Research the outcomes of a mid-course review that assessed progress and considered options for the future of these projects, including the recommendations from the External Consultants Panel (ECP) for the modENCODE and ENCODE Projects that conducted the review. At that time, the Council approved the ECP's recommendation to extend the production projects for a fifth year (through FY2012) to take advantage of the production capabilities that had been established and to provide NHGRI with additional time to plan the future of the program.

In addition to the ECP recommendations, the following proposals were based on additional input from the planning meeting "Genomics of Gene Regulation" held in October 2009 (http://www.genome.gov/Pages/About/Planning/October2009_GenomicsGeneRegulation.pdf) and from the NHGRI Strategic Planning meeting in July 2010 (http://www.genome.gov/Pages/About/Planning/2011NHGRIStrategicPlan.pdf). The gist of the collective input from these several sources is that the ENCODE and modENCODE Projects have been very successful in generating large amounts of high-quality data that are made rapidly available to, and are being heavily used by, the research community, with the result that there has been a significant increase in the understanding of genome function. (See: Nature 447:799-816, 2007; Science 330:1775-1787, 2010; Science 330:1787-1797, 2010; PLoS Biology 9(4): e1001046, 2011.) However, by the conclusion of the five-year full-scale production period, these projects will have interrogated only a fraction of the cells and tissues needed for a comprehensive catalog of functional elements. Generation of truly comprehensive catalogs will require revolutionary new technologies to dramatically reduce the cost and increase the sensitivity of finding functional elements, and applications to develop such technologies are being sought through the above-mentioned technology development initiative. However, through this Concept Clearance, NHGRI is also proposing to continue focused production-scale activities directed at expanding the catalog of functional elements using state-of-the-art methods because there is considerable value in supporting simultaneous and collaborative production and technology development efforts to drive the development and implementation of robust methods.

Proposed Research Scope and Objectives:

RFA: Expanding the Encyclopedia of DNA Elements (ENCODE) in the Human and Model Organisms

This RFA would include requests for proposals to:

A. continue work on the annotation of the human genome by improving gene models on the basis of new data on RNA transcripts coming from next-generation sequencing platforms;
B. annotate the mouse genome similarly;
C. expand the repertoire of data types in ENCODE, particularly more classes of RNA molecules and functional elements within RNA molecules for the human and mouse genomes; and
D. take existing data sets more deeply in the human genome (with limited studies in ENCODE model organisms, i.e., mouse, fly and worm). For most data sets, this will involve interrogation of additional cell types; for mapping sites of transcription factor binding, this will focus primarily on greatly expanding the number of transcription factors studied. Because over 1400 transcription factors have been identified in the human genome, and many exhibit tissue-restricted expression, the number of cell types per factor studied may need to be limited. Specific areas for additional studies include:
    i. mapping binding sites for all transcription factors, using at least two cell types for each new factor;
    ii. mapping sites of open chromatin in more cell types;
    iii. mapping selected histone marks and other relevant chromatin proteins in more cell types; and
    iv. mapping sites of DNA methylation in more cell types.

NHGRI seeks to capitalize on the progress that has been made in establishing high-throughput and efficient production pipelines by supporting ENCODE data production centers that will take advantage of inherent economies of scale as well as more centralized management and coordination (e.g., standardized cell sources and data formats). Each center could be focused on utilizing one or several high-throughput methods. The primary focus will be to support work to further the catalog of functional elements in the human genome and the secondary focus will be to support efforts in the mouse, particularly where noted above.

This RFA would also allow the submission of proposals to support continued studies in *C. elegans* and *D. melanogaster*, but such projects will comprise a smaller fraction of the ENCODE production effort going forward. Much progress has been made in cataloging functional elements in these two model organisms, and although their catalogs are not complete, NHGRI has concluded that the science has progressed to the point where a large, centralized effort is probably not needed. The increased accessibility to high-throughput sequencing technologies, which are at the core of many of the methods being used by these projects, has changed the community's ability to generate genome-wide data rapidly and efficiently to study specific biological questions. The modENCODE Project has provided a firm foundation for these more focused studies, and the consortium infrastructure is, therefore, no longer essential in NHGRI's opinion.

With the vast amount of data that the modENCODE projects have already generated, NHGRI plans to take advantage of these datasets and the utility of these model organisms to begin to address the important questions of how functional elements in the genome interact with each other, how they assemble into networks, and how genome function can be predicted from primary sequence information. (See separate concept clearance on Genomics of Gene Regulation.)

To help ensure that the quality of the data is high and that the catalogs are useful to the research community, each center will be allowed to include limited, well-chosen biologically-focused projects to take advantage of their established pipelines to perform biological experiments for further characterization and validation of the ENCODE data.

Mechanism of Support: These projects will be supported by cooperative agreements using the U54 Center mechanism. The amount of funds devoted to this RFA will be approximately $15-25 million total costs per year, for four years. It is anticipated that 6-8 awards will be made.

RFA: A Data Analysis and Coordination Center for the Encyclopedia of DNA Elements (ENCODE) Project (DACC)

The purpose of this RFA is to solicit applications to implement a Data Analysis and Coordination Center (DACC) to serve as a centralized database for the ENCODE Projects and to coordinate the analysis of the ENCODE Projects data. This represents a consolidation of four separate activities currently supported under ENCODE and modENCODE (i.e., a data coordination center and data analysis center for each consortium) that should result in operational efficiencies. The DACC will be funded to develop, house, and maintain databases to track, store, and provide access to the data generated by the ENCODE Projects.

The DACC will also provide an informatics resource to ensure consistent data analysis and to facilitate the integrative analyses of the various data types being generated from multiple platforms being studied in ENCODE. One key aspect of these analyses will be to define a minimum set of elements/marks needed to identify a unique molecular signature of the cell to optimize the data generated by ENCODE and other related projects. The responsibility for the integrative analyses of the ENCODE data will continue to rest with the ENCODE Analysis Working Group (AWG), as will responsibility for ensuring the maximum utility of the ENCODE data for the community. (The AWG will be comprised of relevant members of each of the production groups, each of the analysis groups supported under the companion analysis RFA, the DACC itself, and any additional analysis groups that join the ENCODE Consortium.) The DACC will work with the AWG to identify the types of analyses that need to be performed and to perform all necessary data transformations to enable these analyses. The activities of the DACC will be complementary to, and not overlapping with, the

local informatics activities of each production center. Thus, the DACC will be part of an essential annotation resource to ensure that the results of the ENCODE Projects are made useful for the scientific community.

Mechanism of Support: The DACC will be supported through the U41 Cooperative Agreement mechanism. $3.5 million total costs per year, for four years, will be set aside for this initiative. It is anticipated that one award will be made.

RFA: Computational Analysis of the Encyclopedia of DNA Elements (ENCODE) Data

ENCODE will provide a reference catalog of functional elements that the community can use to expand on basic biological knowledge and to interpret disease mapping studies. To enhance these activities, NHGRI plans to solicit applications from researchers outside of the umbrella of the ENCODE Project to support analysis activities on the ENCODE data. These activities might include combining ENCODE data with related functional genomic data to derive new biological insights, using the ENCODE data to improve on the analysis of disease mapping studies to identify causal variants, or developing new methods to improve on analysis and interpretation of ENCODE data. It is expected that the awardees will participate in the ENCODE AWG activities.

Mechanism of Support: Projects to enhance the analysis of ENCODE data will be supported through the U01 Research Project grant mechanism. $3 million total costs per year, for three years, will be set aside for this initiative. It is anticipated that 6-10 awards will be made.