National Advisory Council for Human Genome Research
7 February 2011

Concept Clearance for RFA

Technology Development for High-Throughput Functional Genomics

Background: The long-term goal of the Encyclopedia of DNA Elements (ENCODE) Project is to identify all of the sequence-based functional elements in the human genome. modENCODE is a parallel effort to annotate the genomes of *C. elegans* and *D. melanogaster* comprehensively. Since their inception, ENCODE and modENCODE have generated very large datasets from which a great deal of functional information has been extracted. Very importantly, integrating the analysis of datasets within and across a variety of cell types has amplified the amount of useful information that has been obtained from these datasets (Science 330:1775-1787, 2010; Science 330:1787-1797, 2010). The use of multiple cell types is, in fact, key to obtaining a comprehensive description of the functional elements in these genomes. And, while ENCODE and modENCODE are on track to meet the goals as currently described, those goals still fall short of achieving a truly comprehensive view of the structure of the functional elements in these important genomes.

The development and use of new technologies, some of which were developed earlier in the ENCODE project, have been important in bringing the ENCODE and modENCODE programs to their present state. However, revolutionary new technologies will be needed if they are to reach the goal of comprehensiveness, particularly in the case of ENCODE with the challenges posed by the much more complex genome sequence, the number of cell types and the accessibility of tissues for study. For example, the number of cells currently required for the ENCODE assays (in the tens of thousands) means that most of the results to date come from cell lines, which may or may not adequately represent *in vivo* conditions. Some work has also been done with tissues, but the mixed cellular nature of most tissues prevents analysis of relatively rare cells in mixed populations or cell populations from the human body where the quantity of tissue samples might be very small. In either case, current technology is very far from allowing the ideal of single cell analysis. Furthermore, the relatively high cost of the technologies now available limits the number of combinations of cell types and assays for functional elements that can be tested, as well as the application of these technologies to disease studies.

The data produced by ENCODE (and the parallel modENCODE Project) are critical to advancing further research in understanding the “rules” by which genome sequences are “read” and genetic information is expressed. This was the subject of an NHGRI workshop on the “Genomics of Gene Regulation” held in October 2009 (http://www.genome.gov/Pages/About/Planning/October2009_GenomicsGeneRegulation.pdf). Attendees at this meeting were another group that strongly supported the value of these catalogs and recognized that more technology development was

needed to complete them. They also noted that complete ENCODE–type data will be key in attempts to understand how genetic variation gives rise to different phenotypes. Thus, one of the major recommendations from the workshop was that support is also needed for the development of better experimental approaches to model gene regulation at a genome scale. For example, many of the current methods used in ENCODE identify putative functional elements through the use of surrogate marks such as histone modifications associated with particular specific biological activities. High-throughput methods are needed to validate these inferred biological roles (e.g., enhancers, insulators, splicing regulators) to add value to the ENCODE datasets and to enable tests of predictive models that describe how functional elements interact to drive gene expression.

Proposed Research Scope and Objectives: NHGRI proposes to solicit applications through the issuance of two RFAs, for research projects to develop new technologies that go well beyond currently available methods and can be used to comprehensively identify and biologically validate sequence-based functional elements of a given type in eukaryotic genomes, with particular emphasis on methods that will enable annotation of the human genome sequence. The intent of this solicitation is to support the development of novel, robust approaches to: 1) reduce the cost of identifying functional elements by several orders of magnitude; 2) reduce the amount of biological material required to identify functional elements, moving towards single-cell analysis; and 3) increase the throughput of biological validation assays.

Mechanism of Support: The mechanisms of support will be the R01 research project and the R21 pilot project. The total project period for applications submitted in response to these RFAs may be three years for R01s and two years for R21s. NHGRI plans to set aside $5 million per year for three years for these RFAs with a maximum of $350,000 direct costs per year for R01 applications and $275,000 direct costs over two years for R21 applications.