National Advisory Council for Human Genome Research

May 16, 2011

Concept Clearance for RFA

Demonstration Projects for the Genomics of Gene Regulation

This is a Concept Clearance to initiate a new project, the Genomics of Gene Regulation, which is intended to explore new approaches to the investigation of how genome sequence dictates gene expression. The proposed RFA is to solicit demonstration projects to launch this initiative.

Background: One of the great challenges to interpreting the human genome sequence is to identify regulatory sequences and to understand how they control where and when genes are switched on and off, and their level of expression, during development and in response to environmental stimuli. To begin to address this challenge, the NHGRI established the Encyclopedia of DNA Elements (ENCODE) Project to identify all of the sequence-based functional elements in the human genome. modENCODE, a parallel effort to annotate the genomes of *C. elegans* and *D. melanogaster,* set a similar goal with the expectation that insights into these genomes would help to establish general principles that could be applicable to the human genome as well as facilitate an understanding of the biology of these two important experimental models. In 2009, with funds from the American Recovery and Reinvestment Act (ARRA), a small, two-year effort to generate similar data from the mouse genome was established to allow additional comparative analysis and to aid in the evaluation of the quality of the human ENCODE data. Since their inception, ENCODE and modENCODE have generated very large datasets from which a great deal of functional information has been extracted. Significantly, 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; PLoS Biol 9:e1001046, 2011).

These catalogs of functional elements provide a rich annotation of the human, fly and worm (and to a lesser extent the mouse) genomes and are increasingly being used by researchers around the world who are interested in gaining a deeper understanding of gene regulation, either globally or on a locus-specific basis, as well as by disease researchers interested in deciphering the meaning of GWAS hits, especially those falling in non-coding regions of the genome. A next step in determining the role of functional elements in regulating gene expression is to understand how particular elements interact with each other to give rise to specific gene expression signatures in sufficient depth to be able to decipher the "language" of gene regulation. With an adequate understanding of the rules under which gene regulation operates, it should be possible to "read" DNA

sequence and predict where, when, and at what level any gene will be expressed. One approach to obtaining such an understanding is to develop predictions to describe how sets of elements interact to generate a gene expression signature, experimentally test these predictions and iteratively develop new predictive computational models as additional data are available.

To explore this approach, NHGRI proposes to initiate a new program, the Genomics of Gene Regulation (GGR), starting with the funding of demonstration projects to explore the feasibility of learning the rules by which genes are regulated, and of exploiting these rules to predict gene expression signatures in a given cell or in particular stage of development. The proposed GGR demonstration projects will provide an opportunity to leverage existing ENCODE, modENCODE and related functional genomic data as a useful start to this process.

Proposed Research Scope and Objectives:

The long-term goal of GGR is to apply genomic approaches and to use genomic technologies to understand how genetic regulatory systems are assembled and then function to determine biological processes at a mechanistic level. While the architectures of a small number of complicated genetic regulatory systems have previously been determined by using the techniques and approaches of genetics and molecular biology, the development of comprehensive catalogs of functional elements in the genome of an organism and of genomic technologies provide new and powerful tools for dissecting multi-component genetic regulatory circuits and understanding the underlying "grammar" that cells use to assemble individual systems from the large set of functional elements encoded in the genome of an organism. We propose that the GGR program begin by funding a set of projects that attempt to demonstrate that genomic data and genomic technologies can be used to advance the understanding of genetic regulatory systems, where understanding is defined as being able to build a predictive model that can be experimentally tested and refined.

A proposal responsive to this RFA will describe a project that will use genomic approaches for the comprehensive identification of the components and organization of a genetic regulatory circuit in a well-established experimental model system. The project should be able to take advantage of the data produced by the ENCODE and/or modENCODE Projects, if possible, and use those data in the generation of hypotheses about the details of regulation that are described by an initial gene regulation model. Such a model should describe the biological role of the identified functional elements and how they interact to form a genetic regulatory circuit. The predictions of the initial model would then be tested experimentally by appropriate assays, for instance by perturbing (by genetic or environmental means) identified key regulators and comparing the measured biological response to the predicted response. Analysis of the outcomes of these experiments would be used to refine and further develop the

model in an iterative process of modeling, prediction, and experimentation. It is hoped that this approach will lead to a detailed understanding of the individual genetic regulatory circuits under study. It is further hoped that by investigating several genetic regulatory circuits this way and organizing the projects in a collaborative research network, the GGR investigators will be able to decipher the grammar rules that determine genetic regulation, to identify if there are rules that are generalizable, and to establish the methodology to expand this approach to other regulatory pathways.

Although each funded project will likely focus on a different biological system, awardees will be expected to interact with each other in a Research Network to enhance information exchange.

Each demonstration project would be expected to have the following goals:

1) Comprehensive identification of the functionally active elements and the gene expression signatures present in a ground cell state and, for comparison, in a cell state that is induced either by normal development or by an environmental stimulus.

2) Generation of a predictive model that describes the biological role of the identified functional elements and how they interact to establish the gene expression signatures observed in the ground and induced states.

3) Experimental testing of the predictions that the model makes about the biological role of the functional elements and their interactions. In some cases, these experiments may involve testing the effects of perturbation of individual components of the predicted regulatory circuit on the gene expression signature.

4) Iterative refinement of the predictive model based on experimental results.

5) Comparison of the experimental results and validated model with those developed by other participants in the GGR Research Network with the goal of developing a description of the grammar of gene regulation.

Mechanism of Support: The primary mechanism of support will be the U01 cooperative agreement. $10 million total costs per year for 3 years will be set aside for this initiative. It is anticipated that 5-8 cooperative agreements will be awarded.

Related Activities: To facilitate data dissemination, these projects will submit data to the ENCODE Data Analysis and Coordination Center (DACC) that will be funded as part of the next phase for the ENCODE Project. To further stimulate

research in the computational analysis of these data, NHGRI will also issue a Program Announcement to solicit R01 applications for additional analytical projects. Applications received in response to this Program Announcement will be funded from competitive NHGRI funds.