ENCODE
Encyclopedia of DNA Elements

Prospective Applicants Meeting
December 18, 2006

Fishers Lane Conference Center
Rockville, Maryland

Challenge

Compile a *comprehensive encyclopedia* of all of the sequence features in the human genome.

Approach:

- Apply lessons learned from the success of the Human Genome Project
- Start with well-defined pilot project
- Develop and test high-throughput technologies

Phased Approach to ENCODE

Phase 1: Pilot Project using Existing Technologies

Research Consortium focused on identification of transcription units, transcriptional regulatory sequences, DNase hypersensitive sites, chromatin modifications, and origins of replication

Phase 2: Technology Development

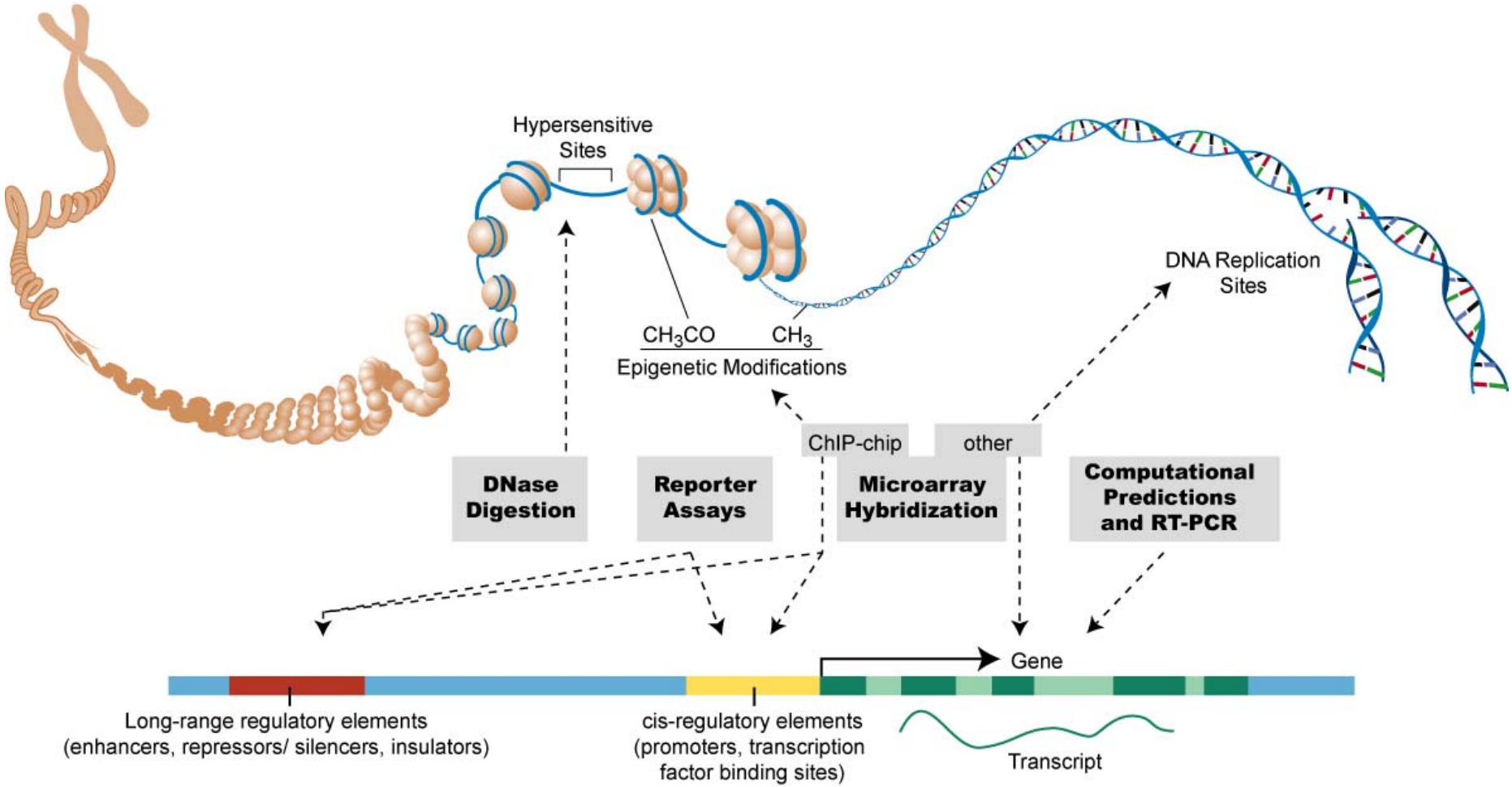
Focused on less well-studied functional elements

Phase 3: Expanded Pilot Project

Solicit new applications

ENCODE Pilot Project Goals

- Test and compare existing and new methods for exhaustive identification and validation of functional sequence elements in a limited region of DNA (~1% = 30 Mb)
- Identify gaps in ability to annotate genome
- Set a clear path for scaling up this effort to efficiently and effectively characterize the entire human genome in detail


Hypersensitive Sites
DNA Replication Sites
CH_3CO
CH_3
Epigenetic Modifications
ChIP-chip
other
DNase Digestion
Reporter Assays
Microarray Hybridization
Computational Predictions and RT-PCR
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

ENCODE Consortium

- Bring together investigators with diverse backgrounds and expertise to work cooperatively
- Iterative process with close interactions between computational and experimental work
- Share resources and expertise within the Consortium
- Work cooperatively to solve problems encountered during the project
- Open to all academic, government and private sector scientists interested in the goals of ENCODE project
- Open to participants not funded through RFA
- Common databases
- Common data release policy

Criteria for Participation in ENCODE Consortium

- Willingness to analyze entire set of target regions (next phase expand to whole genome)
- Offer substantial contribution
- Share all results according to Consortium Data Release Policy
- Participate in group activities
- Funding by NHGRI not required
- Membership Approved by NHGRI staff and External Consultant Panel

ENCODE Consortium Data Release Policy

As a community resource, NHGRI encouraged a data release policy for ENCODE that implements the principle and achieves the advantages of rapid pre-publication data release.

Data Release Policy

1. Participants submit data to the Consortium database(s) as soon as the data have been determined to be reliable.

2. Participants submit data to the Consortium databases in the specified format.

3. To achieve maximal utility of this community resource, all data would be made freely available to the entire research community in a form that would allow for redisplay and reanalysis. Users of these data should respect the legitimate interests of the producers to analyze and publish their results by treating the data as unpublished information, until otherwise indicated.

Data Release Policy

4. Individual investigators within the Consortium may publish the results of their own work.

5. The Consortium will publish global analyses of the pilot project's results in a timely manner.

6. Publicly funded Consortium participants will fully disclose algorithms, software source code, and experimental methods to the other members of the Consortium for purposes of scientific evaluation and will be strongly encouraged to make them available to the broad research community.

Next Phase of ENCODE

- Support efforts to apply high-throughput methods to develop a comprehensive catalog of functional elements in the human genome sequence (RFA HG 07-030)
- Support a data coordination center to house and maintain the ENCODE data (RFA HG 07-031)
- Support similar efforts in selected model organisms (modENCODE) (RFAs closed – funding March 2007)
- Continue to support technology development efforts (RFAs closed – funding June 2007)

Planning for next phase of ENCODE

- Goal of pilot project to set path for scaling to whole genome studies
- Consulted advisors to give feedback on current pilot project and future of ENCODE
  - Consider progress that has been made in the pilot project and current state-of-the-art experimental capability of technologies to conduct comprehensive, genome-wide studies to identify various functional elements in the human genome
  - Concluded that pilot project going very well and planning for the next phase should continue
  - Made a number of recommendations for considering scaling projects

Key points for next phase of ENCODE

- Open competition to allow for new ideas/participants
- Allow for continued work on 1%
- Support efforts to scale to the whole genome

Factors to consider in scaling ENCODE to the whole genome

- Scale when data produced would be cheaper and of better quality than what can be done on an individual laboratory basis
- Applicants need to demonstrate a realistic awareness of issues involved in scaling
- Technologies should not be in “flux”
- Applicants need to demonstrate a good understanding of quantitative metrics of data quality
- Validation pipelines need to be in place
- Robustness of the technology must be demonstrated
  - Sensitivity
  - Specificity
  - Biological validation
- Biological utility of the data should be high
- Costs should be reasonable and well-documented

RFA: Creating the Encyclopedia of DNA Elements (ENCODE) in the Human Genome (HG 07-030)

Primary Goal: Comprehensive identification of functional elements in the human genome sequence using high-throughput, cost-effective and highly sensitive methods

Support two types of efforts:

Whole-genome scale

Pilot project scale (30Mb target regions)

- Continuation of on-going studies
- New studies

Creating the Encyclopedia of DNA Elements in the Human Genome

- Use U54 and U01 cooperative agreement award mechanisms
- $23M set aside in FY07
- Anticipate funding 6-10 awards

- Scientific focus on finding and validating functional elements, for example:
  - Transcribed sequences
  - Conserved non-coding sequences that specify functional elements
  - Cis-acting DNA elements that regulate transcription and/or chromatin states, e.g., promoters, enhancers, repressors, insulators
  - DNA sequence features that affect and/or control chromosome biology, e.g., origins of replication, hot spots for recombination
  - Epigenetic marks, e.g., DNA methylation, chromatin modifications

Creating the Encyclopedia of DNA Elements the Human Genome

- Primary purpose of RFA is to generate high-quality, comprehensive catalog of functional elements for the community to use to answer biological questions
  - Applicants should not plan to address specific biological question
- Use of centralized, genome-scale approach should result in advantages such as cost efficiencies and higher data quality over smaller scale efforts in multiple laboratories = seeking “bang for the buck”

Data Quality and Comprehensiveness

- Describe plans to verify that data is reproducible
- Describe plans to validate biochemical authenticity of data using preferable completely independent method on subset of data to provide 95% confidence level that events occur in vivo
- Describe plans to provide information on the biological relevance of the data, i.e., confirming the biological authenticity of the class of elements being identified using a subset of data to provide 95% confidence that the elements do have the inferred function in vivo
  - These studies are limited to no more than $300,000 direct costs/year
- Define “comprehensiveness” expected for project, describe ability to identify significant fraction of functional elements studied and provide estimate of sensitivity of proposed approach

Production Issues

- Operating at Scale
  - Provide evidence of current capacity, how production goals are set and met, any plans for further pipeline development
  - May ramp-up in 1-2 yrs
- Pipeline Description
  - Describe individual steps/components of process and how they are integrated, resources needed for each step
- Quality Control
  - Describe QC steps in production process, process for determining failure rates, ensuring high-quality product
- Costs
  - Demonstrate understanding of costs and how to track them; develop cost model, considering all aspects of the pipeline; describe anticipated cost reductions through project period
- Milestones and Goals
  - Set goals for overall project and annual production milestones with metrics that will document progress toward achieving goals
  - Milestones may be negotiated at the time of the award

Bioinformatics Issues

- Discuss all pertinent informatics issues associated with defining functional elements
  - Describe informatics pipeline for processing primary data to generate a list of functional elements
  - Include informatics and statistical tools needed to integrate primary data and any validation data obtained using different method(s)
  - Limit integrative analyses to those needed to cross-validate conclusions from each technology
  - More in-depth analyses of data are outside of scope of RFA

Whole Genome Studies

- Sufficient preliminary data to support application of method to entire human genome sequence (e.g., demonstration of ability to apply method at high-throughput with high (known) data quality and ability to approach “comprehensiveness”)
- Technical approach should be “hardened”
- Apply method to entire human genome sequence
- Need to fully address production pipeline issues
- May include “ramp-up” of up to 50% in first 1-2 years
- Use U54 Research Center award mechanism
- Research Plan page limit: 40 pages
- Up to 4 years of support
- P.I. must devote a minimum of 25% effort

Pilot Projects

- Proof of principle already established, but insufficient preliminary data to support application of method to entire human genome sequence (e.g., demonstration of ability to apply method at high-throughput with high (known) data quality and ability to approach “comprehensiveness”)
- Technical approach may still be under flux
- Focus exclusively on 30 Mb target regions
  - May include use of whole-genome approaches with validation and analysis focused on target regions
- Encouraged to discuss production pipeline issues as appropriate
- Use U01 Research Project award mechanism
- Research Plan page limit: 25 pages
- Up to three years of support
- No minimum effort for P.I. required

Clarification on Antibody Production

- In RFA, states that “generation of a large set of antibodies to DNA binding proteins for e.g., ChIP-chip studies, although valuable, will be considered outside of scope of this project. NHGRI is considering other means to support such an effort.”

- CLARIFICTION: Applicants can propose to generate antibodies needed to support experiments in application. As NHGRI’s plans are developed, we may work with funded groups to ensure that there is no redundancy of antibody generation.

Timeline for RFAs HG 07-030 and HG 07-031

Letter of Intent Due	February 27, 2007
Application Receipt Date	March 29, 2007
Peer Review Date	June/July 2007
Council Review Date	September 2007
Anticipated Funding Date	September 30, 2007
Consortium Meeting	November 28-29, 2007

General Guidance

- Read the RFA very carefully and completely
- Formulate ideas and questions
- Contact Program Staff as early on as possible
  - Send ideas and questions by email first
  - Do not need to wait until Letter of Intent Deadline
- In proposal, address each of the 12 points summarized at end of Research Scope