Report From A Planning Workshop
for the National Human Genome Research Institute
"From Genome Function to Biomedical Insight: ENCODE and Beyond"
March 10-11, 2015

Executive Summary

On March 10-11, 2015, the National Human Genome Research Institute (NHGRI) convened a workshop to discuss scientific questions and opportunities that can be addressed by large-scale functional genomics studies, and options for future NHGRI programs in this area following the conclusion of the current phase of ENCODE. ENCODE (Encyclopedia of DNA Elements) aims to develop a comprehensive map of functional elements in the human and mouse genomes. The workshop was video cast and archived (see http://www.genome.gov/27560819).

The workshop was organized around three scientific opportunities that would use genomic assays of function to: identify and characterize functional elements in an unbiased manner; interpret disease-associated genetic variation; and address basic biological questions. For each, illustrations of projects that could address unmet scientific needs were presented, followed by discussion of key themes. The workshop concluded with a discussion of goals and priorities spanning the three topics, including discussion of implementation (see Appendix 1: Agenda).

There was strong support for pursuit of all three opportunities. Participants were enthusiastic about ENCODE accomplishments to date and the opportunities that flow from that resource. Important themes included that NHGRI should: focus on catalytic work such as developing resources and new approaches; leverage the expertise and efforts of individual scientific leaders in the relevant areas of human disease and biology; and continue to fund large-scale functional genomics science. These foundational activities would support the main conclusions of the workshop, which can be summarized as strong support for:

- Using genomics assays for genome-wide identification of functional elements
- Adding functional characterization efforts
- Applying functional genomics assays directly to disease and biological studies
- Increasing community participation
    - Community samples
    - Community data

This work is required to interpret genetic studies of human disease.

In summary, the group agreed the highest priority for future work is to expand beyond cataloging genomic elements to understanding the role of these elements in genome function. This can be achieved by continued mapping efforts, a new functional characterization effort, increased community participation, and direct disease studies. This combination is expected to enhance genetic studies of human disease.

Report From A Planning Workshop
for the National Human Genome Research Institute
"From Genome Function to Biomedical Insight: ENCODE and Beyond"
March 10-11, 2015

On March 10-11, 2015, the National Human Genome Research Institute (NHGRI) convened a workshop, the objectives of which were (1) to discuss scientific questions and opportunities that can be substantively addressed by large-scale functional genomics studies, and (2) to consider options for future NHGRI programs in this area following the conclusion of the current phase of the ENCODE project in 2016. ENCODE (Encyclopedia of DNA Elements) aims to develop a comprehensive map of functional elements in the human and mouse genomes. NHGRI staff and the organizing committee (Carol Bult, Eric Boerwinkle, John Lis, Aviv Regev) developed the workshop in a collaborative manner. The workshop was video cast and archived; materials are freely available at http://www.genome.gov/27560819.

The workshop began with introductory material including overviews of functional genomics research, progress in the current phase of ENCODE, and reports from related activities. The core of the workshop discussion was organized around three scientific opportunities: first, using genomic assays of function to identify and characterize functional elements in an unbiased manner; second, using genomic assays of function to interpret disease-associated genetic variation; and third, using genomic assays of function to address basic biological questions. For each topic, illustrations of projects that could address unmet scientific needs were presented followed by discussion of themes that emerged from those presentations. The workshop concluded with a discussion of goals and priorities spanning the three topics, including some discussion of implementation (see Appendix 1: Agenda).

Background Presentations and Discussion

Eric Green, director of NHGRI, introduced ENCODE as a project that was built on the Human Genome Project, an important part of the 2011 NHGRI strategic plan, and a vital part of the institute's focus. He explained that the ENCODE project plays to NHGRI strengths, as a technology-driven consortium project, and as a resource- and hypothesis-generating project, without focusing on a particular biological system or disease. ENCODE is relevant to understanding the genetic basis of disease and to precision medicine. Elise Feingold, program director at NHGRI, reinforced the director's comments about the importance of functional genomics for the mission of NHGRI. The goals, history, approach, and structure of the ENCODE project were briefly presented, as were highlights of ENCODE's accomplishments.

The NHGRI goal for the workshop was to hear thoughts from representatives of the biomedical research community about functional genomics, in order to assess the scientific landscape, identify gaps, and consider future priorities. NHGRI will incorporate feedback from this workshop in programmatic discussions, in the

context of the 2011 NHGRI Strategic Plan and the Extramural Research Program, to present a proposal for future projects for consideration at the May 2015 National Advisory Council for Human Genome Research (NACHGR) meeting.

The workshop participants spanned a breadth of scientific expertise and community roles. Representation included technology developers, as well as researchers studying particular biological systems and diseases; independent investigators and members of various consortia; NHGRI council members and ENCODE External Consultants Panel members; ENCODE data users and ENCODE consortium members; as well as extramural, international, and NIH intramural scientists (see also Appendix 2: Participant List).

Presentations:

Ewan Birney provided a functional genomics overview by describing the impact of ENCODE data on understanding biology and human disease, how ENCODE data are used, and projected community needs over the next decade. Dr. Birney indicated that the resource enabled scientific exploration and discovery, through sharing ENCODE data and analyses in an unrestricted manner. The presentation highlighted the large number of community publications using the ENCODE resource, and the value, for informing a wider community about the utility of the resource, of sharing the list of these publications. Examples were presented of how the resources facilitate research in basic biology, common disease, and cancer. He suggested that systematically collected data results in higher value to the community than aggregated data, and there is compelling need for more data for both regulatory elements and DNA-DNA interactions. Model organisms could be especially valuable for testing the role of candidate functional elements in organismal phenotypes. Finally, the value of predictive, modeling-based approaches was presented as a needed advance to interpret the functional consequences of genetic variation.

BOX 1: Key points emphasized during Background Presentations and Discussion

- Current ENCODE resource is useful for studies of biology and disease
- Current ENCODE resource has been used in hundreds of publications by consortia members and the broader scientific community
- Users require high-quality data, collected in a systematic manner
- Systematic mapping in diverse cell fates and cell states continues to be of high value
- Profiling cell fates and cell states that are especially relevant to disease would be of high value
- Predictions of functional connections between regulatory elements and genes should be enhanced
- Effective mechanisms are needed to efficiently share transcriptomic and epigenomic data
- A systems biology approach is needed to advance the study of genome function

Next, to provide participants with an equal footing in their knowledge of current functional genomics efforts, a survey of large functional genomics projects was presented. An overview of ENCODE goals, approaches, and accomplishments was presented by Mike Snyder (see also Appendix 3: ENCODE Summary). Brief overviews of other related projects funded by U.S. and international agencies were presented by Daniel Gilchrist and Mike Pazin, along with some analysis of the defining features of these projects and their interactions with ENCODE (see also Appendix 4: Projects With Similarities To ENCODE).

Recommendations for future projects that were made by current ENCODE PIs were presented by Joseph Ecker (see also Appendix 5: "ENCODE 2020: From Elements to Function"). Their theme was that NHGRI could make enabling contributions to functional genomics by focusing on areas where the coordinated action of a consortium and large-scale data generation can have the most impact. First, there is considerable value to continued comprehensive mapping of functional elements in an unbiased manner. This work should incorporate new mapping tools, as well as samples and data provided by the broader community. Second, using a combination of experimental and computational techniques to connect functional elements to their cognate genes could enable correct identification of the genes and pathways that are relevant to diagnosis and treatment of disease. Third, it is important to identify what functions (if any) are performed by candidate elements and how they work (e.g., increase RNA production by acting as boundary elements, or alter exon inclusion by acting as splice-site enhancers). These functional characterization activities are beyond the scope of the current ENCODE project. Fourth, to understand the role of individual genetic variation in human disease, individual variation in sequence elements and its impact on quantitative phenotypes and disease must be studied. This is especially true of non-coding variation (the vast majority of common variants associated with disease lie outside of protein-coding regions).

Recommendations from the NHGRI Genome Sequencing Program workshop (July 28-29, 2014) break-out session on genome function (see also http://www.genome.gov/27558042) were presented by Mark Gerstein and Richard Myers, co-leaders of the session at the July 2014 workshop. It was suggested that there is significant value both to catalog genomic functional elements in an unbiased manner, and to characterize candidate genomic regions apparently associated with diseases or traits of interest. Defining the function (if any) of both coding and non-coding candidate variants was regarded as foundational for genomics. While many functional genomics projects assay entire genomes (including coding and non-coding sequences), adding the dimension of variation within populations, and the dimension of variation over time, would bring the field closer to realizing its full potential. Finally, some workshop participants voiced the opinion that function should be considered at the systems biology or network level, because an understanding of how the parts work together would lead to greater insight. (The sequencing workshop report is available at: http://www.genome.gov/27559219).

Discussion:

During discussion there was agreement that generating and sharing predictions of linkage between regulatory elements and genes would have great value. As the best way to make these predictions is not currently known, the evidence base used in generating such predictions must be transparent. Some participants indicated they were already using existing ENCODE predictions in their work. Future work could include experimental characterization of these predictions, and collaboration with the Genotype Tissue Expression (GTEx) project, which is also making these kinds of predictions using a different approach based on different data. There was also enthusiasm for systems-based approaches, such as predictions of the effects of genetic variants on gene expression, and consideration that genetic variants (both within and across genes) can interact in non-linear ways, and this should be modeled.

There was agreement that it is very important to efficiently and transparently share data; varying opinions were expressed on how to achieve this goal. Some supported splitting the ENCODE DCC data broker function from the genome browser function; a strength of the current DCC is focusing on presenting data and metadata in a manner that makes it interoperable with other projects, which scales better than trying to create a database of all of human biology. Others preferred the more traditional browser-centric approach to interacting with the data, and suggested that NHGRI consider a visualization portal for all genomic data. It would help the community if data from multiple consortia were collected at one location; it would be even more valuable if these data from different projects were uniformly processed.

The value of transcription factor mapping data for disease studies through enabling predictions of regulatory machinery that is upstream of the genomic elements was endorsed. There was recognition that the value of existing ENCODE data is derived in part from the high quality of the data, and in part from the systematic manner of data collection, and these should be important features of any future project. There was strong agreement that continued, unbiased cataloging is highly useful. These efforts should be expanded to include profiling cells in relevant cell states (stimulated, disease, etc.).

Topic 1: Identifying and characterizing functional elements

Presentations:

Topic 1 began with presentations from Ross C. Hardison, Laurie Boyer, Frank Pugh, and William Greenleaf. Presenters explained the value of mapping epigenomic and transcriptomic features. They emphasized the value of including 3D interaction maps, and incorporating new technologies to map features and scales not routinely covered today (e.g., assays that measure features with a resolution between 200 and 2000 bp are not widely used today). Assays using small numbers of cells, or even

single cells, were called for to measure signatures that would be lost in ensemble measurements. The importance of experimentally characterizing the function of candidate elements was raised (this was largely outside the scope of current and past ENCODE efforts). Studies incorporating system dynamics and perturbations would improve our understanding of the molecular phenotypes associated with particular functional elements. It was suggested that new assays and analyses combined with functional characterization could expand and refine our vocabulary for regulatory elements.

BOX 2: Key points emphasized during Topic 1 presentations and discussion

- Unbiased mapping of epigenomic and transcriptomic features has high value
- More 3D interaction maps are needed to link regulatory elements to genes
- New assays are needed to map new features
- Single cell/low cell number assays are urgently needed
- Increased characterization and more nuanced understanding of element functions is needed
- Perturbations and dynamic measurements are needed to determine causality
- Dramatic new opportunities arise from new genetic tools (CRISPR)
- Increased community involvement has many benefits
- Data standards, data interoperability, and data visualization tools are important
- Enhancers could be key elements for understanding the cell-specific nature of gene expression

Presenters enthusiastically endorsed the importance of increased community participation in future functional genomics projects. Better data visualization was endorsed to facilitate improved interpretation and broader use of the data. The critical importance of maintaining stringent data standards, data and metadata coordination, and open data sharing for the project were also stressed. It was suggested that future work would require an expanded data coordination center to accept community data and metadata submissions, assess compliance of community data and metadata with standards, and increase interoperability with other projects.

Speakers suggested that today it is easier to identify genetic variation, and to identify candidate variants associated with diseases and traits, than to identify the functional impact, causality, or underlying mechanism; therefore, functional characterization has become a bottleneck that needs to be addressed in order to interpret the role of genetic variants in human disease. Major uses for catalogs of candidate functional elements continue to be prediction of causal variants, pathologic cell types, upstream regulators, and downstream target genes.

Discussion (Carol Bult – Moderator):

It was agreed that functional characterization of the non-coding genome is an important goal. A number of challenges were identified, such as understanding

interactions and cell specificity of elements, and determining the best assays to identify and characterize function of elements. The catalog of elements is very useful in guiding these experiments, by allowing predictions of what function to test, in what cell type; the dimensionality of conditions that require exploration could otherwise be too high. Further study of enhancers is important, because they integrate signaling information in cells, which leads to functional genomic outputs. It was agreed that understanding transcription factors that function upstream of cis-regulatory elements has value in the study of biology and disease. There was enthusiasm for refining our vocabulary for regulatory elements beyond the well-studied concepts of enhancers and promoters, to increase understanding and predictive power. There was also strong enthusiasm for increased community participation, at the level of providing samples to mapping centers, at the level of direct submission of community data to the DCC, and at the level of hearing from disease and biology experts how best to measure biological function. At the end of discussion, the question of how best to investigate function using team science was raised, though no recommendations were made.

Topic 2: Using genomic assays of function to interpret the role of genetic variation in disease

Presentations:

Topic 2 began with presentations from Aravinda Chakravarti, Eric Boerwinkle, Nancy Cox, and William Lowe. While the field has made much progress in identifying candidate genes and candidate variants for human disease, much remains to be done in determining causal variants, causal genes, and mechanisms of disease. A systems biology or network-level understanding of the impact of non-coding elements on gene expression would advance the field, and an understanding of how the parts work together could help to explain how multiple genetic variants can contribute to a particular trait or disease. Statistical power limitations may prevent us from individually mapping most common and rare causal variants associated with disease, unless we incorporate functional data. The value of epigenomic mapping was highlighted in a recent publication that attributes ~80% of the heritability of common diseases to DNase elements identified by ENCODE and the Common Fund Roadmap Epigenomics Mapping Consortium (REMC), and ~8% to protein-coding regions (Gusev et al 2014, PMID: 25439723). Presenters indicated that better tools to predict the effects of genetic variation on gene expression and disease are required. They also provided concrete examples of using functional genomics resources to assign disease heritability to individual variants in two metabolic disorders. Hirschsprung disease was presented as an example where most of the assigned heritability is found in common, non-coding variants that can act additively and combinatorially. Variants associated with this disease appear to affect nearly every aspect of a particular signaling pathway; a systems or network understanding could have guided this work, had that pathway been better understood. Human genetic contributions to disease lie on a continuous spectrum from common/complex, to oligogenic, to rare/Mendelian. To enhance disease

studies, more interactions and sharing are required between the clinical and research enterprises. Today these interactions are often inhibited by privacy rules.

BOX 3: Key points emphasized during Topic 2 presentations and discussion

- Better understanding of the non-coding genome is required, as most of the heritability associated with common diseases lies outside of protein-coding regions
- It is important to map genomic elements in a broad range of cell fates and cell states chosen in an unbiased manner, as well as in cell fates and cell states chosen based on their importance in disease
- Effort should be devoted to functionally characterize a broad range of genomic elements chosen to uncover fundamental principles, as well as genomic elements chosen based on their importance in disease
- Improved categorization of elements and variants is likely to improve the efficiency of functional characterization
- Better predictions of the effects of genetic variation on gene expression will aid disease studies
- Studies of natural variation (genetic and environmental) would add value to gene expression studies
- Model organisms are valuable for testing the role of genomic elements in organismal phenotypes
- Systems approaches are valuable for interpreting the function of genetic variants
- Increased data and sample sharing between the clinical and research enterprises to accelerate disease studies is needed (this is inhibited by privacy rules)

Discussion (Eric Boerwinkle – Moderator):

There was agreement that further mapping studies and functional characterization studies are needed. There were divergent views on whether it was best to perform future mapping studies a) after comparison of the current ENCODE matrix and existing disease knowledge to determine the highest-value cell fates/states that need to be profiled, or b) using unbiased selection of fates/states, as we do not know enough today to identify all the important fates/states. There were also divergent views on whether functional characterization is best done with a) candidate elements associated with diseases, which might be more likely to tell us about human disease, or b) broadly chosen elements and variants (given that it is too early to decide what is important) which might allow us to learn the generalizability of predictions. Indeed, one lesson learned from GWAS is that the candidate disease associations were not necessarily where we would have looked for them (both with respect to the identity of the genes, and the relationship to genes). There was some agreement to recommend both unbiased and disease-driven studies, and that it would be important to obtain some examples from disease to assess the value of this information. About five diseases could be chosen as representative examples to assess the value of these approaches.

There was enthusiasm for utilizing model organisms in testing the role of candidate functional elements in establishing organismal phenotypes,

potentially at scale. There was support for studying natural loss-of-function variants, though views diverged on whether naturally occurring regulatory variants have strong phenotypic effects. More generally, it was seen as important for NHGRI-supported functional genomics projects to begin considering natural variation (beyond disease-associated variation, both genetic and environmental), which was previously considered out of scope. This work could be in collaboration with, or be coordinated with, GTEx. Precise categorization of functional elements identified from mapping studies (and the variants within them) would be a powerful advance, to ensure all categories were characterized, and to ensure the most appropriate assays were used. Finally, proteomics measurements were recommended, for example shotgun proteomics (e.g., starting with transcription factors, or small peptides produced from RNA that was thought to be non-coding).

Topic 3: Using genomic assays of function to study basic biological questions

BOX 4: Key points emphasized during Topic 3 presentations and discussion

- Systematic perturbations are needed to learn the connectivity of genes and regulatory elements
- New assays are required to increase our understanding of genome function
- Developing data standards and uniform data processing approaches for newly emerging assays would have high value
- Functional characterization studies are needed to better understand the existing ENCODE data
- Mapping in new cell types is required to better understand disease
- Incorporation of community samples and data is needed
- A gene-centric view could facilitate use by clinicians
- Some users require a set of standardized assays systematically applied across a wide range of cell types

Presentations:

Topic 3 began with presentations from Brenda Andrews, Karen Adelman, Anjana Rao, and John J. O'Shea. The value of systematic perturbations, network analysis, and how perturbations naturally fed into one type of network analysis was presented. The natural cycle from mechanism, to genomics assay, to mechanistic insight was presented. The value of new assays was emphasized, using as examples assays that measure nascent RNA, and assays that measure DNA modifications beyond methylcytosine. The need for increased incorporation of community samples and community data was again emphasized. Presenters highlighted the need to move from assays correlated with molecular functions, to assays that directly measure molecular and organismal functions. Finally, we heard about the need for a gene-centric view of ENCODE data, and how this could be one tool to help make an individual patient's genetic profile informative to clinicians.

Discussion (Aviv Regev – Moderator):

The group considered how to make functional genomics data useful to the largest audience possible, including clinicians. The importance of data interoperability across projects was recognized as a priority. There was enthusiasm for the current ENCODE DCC efforts to coordinate with worldwide data repositories, to coordinate with other epigenomics projects, and to coordinate with existing ontologies.

The group discussed how to add the expertise of scientists outside of the current ENCODE consortium. There was agreement that it would be good to have experts provide specialized samples to production centers, and it should be possible to reserve a fraction of production capacity for community samples. ImmGen (the Immunological Genome Project; see https://www.immgen.org) was held up as a model where experts willingly collect samples and gene expression data without additional funding, and the aggregated data are of high value to a focused community. There was interest in aggregating existing community data, as well as concern over the expense of acquiring the data, curating the metadata, uniformly processing the data, and performing quality control.

There was enthusiasm for acquiring data from assays that measure new features of established systems, as with new RNA assays and DNA modification assays. Incorporation of new mapping assays could start with a skeletal representation of cell fates/states already profiled, followed by mapping with new assays in the most informative cell types. Strategies for direct fitness-based readout of genetic interactions using pooled cells were considered. The importance of comparing perturbations for cis elements and trans factors was raised. ENCODE could make an important contribution by showing the best way to use new assays, including developing standards and identifying limitations.

Some participants expressed support for profiling many cell fates/states beyond the space that has been explored, noting that beyond the value of identifying new candidate elements, each fate/state profiled reveals new information about the cell specificity of the elements and their importance to disease. Others suggested a deeper understanding of the existing data through functional characterization was a higher priority, and there were likely to be diminishing returns from studying additional fates/states. It was suggested that RNA-seq (at low read depth, and at single cell resolution) could be used to profile large numbers of cell types to learn how much of biology has been sampled, and to identify gaps. The importance of having a systematic data matrix to learn from, in order to make comparisons between samples and assays, was underscored.

Integrative Discussion

During discussion, strong support was expressed for:

- Using genomics assays for genome-wide identification of functional elements
- Adding functional characterization efforts
- Applying functional genomics assays directly to disease and biological studies
- Increasing community participation
    - Community samples
    - Community data

This work is required to interpret genetic studies of human disease.

Discussion (Aviv Regev, Carol Bult and Eric Boerwinkle, Moderators):

A session for integrative discussion facilitated comparison of the themes from the individual topics. Enthusiasm remained high for continued unbiased mapping of candidate functional elements. It was agreed that the most informative assays and cell fates/states remained to be identified, with respect to mapping. Increased community participation was identified as an important component of future work; it was agreed this could happen through community-provided samples and data, and better outreach to make the resource more usable. ENCODE would have to build the infrastructure required to enable this increased community participation. Increased functional characterization of candidate elements was identified as having a central role. This would enable the transition from cataloging elements to understanding how regulatory information is encoded in the genome. Three types of perturbations were recognized as likely to be important in this work: natural genetic variation; engineered genetic variation; and stimuli provided by differentiation, experimental manipulation, or the environment. Well-consented iPS cells could be valuable tools to study natural genetic variation, and provide purified cells for assays. It would be important to have coordination of functional characterization efforts in place at the beginning of the project.

Predictive modeling of gene expression and gene regulatory networks was recommended as an area for future direction. This would provide understanding of genome function, as well as directly facilitate studies attempting to predict the role of genetic variation in non-coding regions on human disease. Modeling should be focused on predicting things that are not easily measured, or predictions that could be used as tests of understanding. The balance between community efforts and consortium efforts remains to be decided, with respect to modeling.

There was enthusiasm for two distinct models for disease studies, conducted in parallel. First, mapping studies that are unbiased (through testing the entire genome) could be performed in cell fates/states that are recognized as especially relevant for disease models, including samples from individuals with disease. Second, functional characterization of candidate disease-relevant variants could be used to connect causal variants to molecular or organismal phenotypes.

Implementation was briefly considered by the workshop participants. There was support for flagship projects, scaled like the current ENCODE effort. Community-provided samples were identified as an important modality to produce data that were relevant to the broader community. Community-provided data was also viewed as an important component. The role of the Data Coordinating Center would have to be expanded to include vetting community data and metadata, and reprocessing community data. Developing and deploying new technologies, along with the necessary data standards and analytics, was also identified as an important component.

Conclusions:

There was broad agreement that it is important to continue functional genomics studies to interpret the role of the genome in biology and human disease. A number of guiding principles emerged:

First, it was agreed that future work should continue to include genome-wide mapping of transcriptomic and epigenomic features. Incorporating new assays was viewed as important to keep the work at the leading edge. The importance of maintaining and enhancing current features of ENCODE, such as rigorous data standards, structured metadata, data visualization and data interoperability, was emphasized.

Second, it was agreed that functional characterization (with a blending of perturbations [genetic and physiological] and mechanistic molecular assays) should be a major component of future work. This is a fundamental departure from past ENCODE efforts, where only limited functional characterization was funded. A large increase in the effort devoted to characterizing genomic elements was viewed as essential to move towards understanding the genome (having an encyclopedia), rather than cataloging the genome (having a parts list).

Third, there was considerable enthusiasm for including direct studies of disease within ENCODE. This would be a significant change from past ENCODE work, which emphasized the study of healthy samples. Two very different approaches were identified as important: 1) Mapping studies performed on samples obtained from diseased and healthy subjects; and 2) Functional characterization performed on candidate elements containing candidate disease-associated variants, with consideration to both low and high risk variants. This would require work in multiple diseases, to attempt to learn what is generalizable about these approaches for studying disease.

Fourth, there was strong support for increased community participation. In part this would be facilitated by use of community-provided samples, which would build on nascent efforts within ENCODE. This would combine the efficiencies of established data collection pipelines, with the knowledge of experts in the biology of

disease. This could also happen by incorporating community-provided data, which would be a new feature for ENCODE. The importance of high-quality data, rigorous data standards, structured metadata, data visualization and data interoperability were all stressed, and it was suggested the data coordination center could be expanded to take on the new role of curating community-provided data. Some of the disease and biology experts might work within functional characterization or computational analysis groups.

Taken together, this work could engage a broader section of the community needed to move from catalogs towards understanding function at the level of the genome, which would support disease studies.

The group did not reach consensus or explore a number of smaller issues. While there was agreement that balance was required between flexibility to introduce new assays versus the value of performing each assay over all cell fates/states in the project, how best to balance these was not decided. This may depend on the type of assay, the scale at which it can be performed (in effort, time, cost), and the amount of existing data from similar techniques. Balance between accepting community samples versus accepting community data was not definitively explored; in particular cases, the deciding factor could well be whether high-quality data already exist for a sample of interest. There was little discussion on the mechanism for prioritizing community-provided samples. The balance between maintaining high data and metadata standards, versus lowering the barrier to submitting community data and metadata, was also not well explored. While there was agreement that it was important to maintain high standards and transparency, there was also concern that they could become barriers to accepting community data. It also remains to be considered whether data submitters would pay the cost of submission, or whether NHGRI would fund a data coordination center to perform this task for data production that was funded by other sources. Finally, the balance between the power of deeply profiling with many assays (with fewer cell fates/states) versus the power profiling many cell fates/states (perhaps with fewer assays) was not agreed upon. Indeed, at the workshop, different types of users outside of ENCODE noted that their work required either many assays in a fate/state, comparison of a broad collection of many fate/state combinations for at least one assay, or a matrix-style data table deep in both assays and cell types.

In summary, the group agreed the highest priority for future work is to expand beyond cataloging genomic elements to understanding the role of these elements in genome function. This can be achieved by continued mapping efforts, a new functional characterization effort, increased community participation, and direct disease studies. This combination is expected to enhance disease studies.

List of Appendices:

From Genome Function to Biomedical Insight: ENCODE and Beyond

March 10-11, 2015
Natcher Conference Center, Rooms E1/E2
National Institutes of Health

Agenda

Objectives:

1. Discuss the scientific questions and opportunities for better understanding genome function and applying that knowledge to basic biological questions and disease studies through large-scale genomics studies.

2. Consider options for future NHGRI projects that would address these questions and opportunities.

Tuesday, March 10, 2015

1:00 p.m.	**Welcome and Setting the Context**	Eric Green
1:15 p.m.	**Purpose of Workshop: Background and Planning Process for Future Initiatives**	Elise Feingold
1:35 p.m.	**From Genome Function to Biomedical Insights: Defining the Scientific Challenges**	Ewan Birney
2:05 p.m.	**Discussion**	
2:30 p.m.	*Break*	
3:00 p.m.	**Genome Function Circa 2016: Updates from Related Projects** Moderator: Daniel Gilchrist	
	ENCODE	Michael Snyder
	REMC/IHEC/BLUEPRINT PsychENCODE Genomics of Gene Regulation (GGR) 4D Nucleome FunVar	Michael Pazin

	FANTOM GTEx LINCS TCGA/ICGC KOMP2/IMPC	Daniel Gilchrist
4:00 p.m.	**Proposals for Future Directions** ENCODE PIs' Vision for Functional Genomics	Joseph Ecker
	Recommendations related to genome function from NHGRI's Planning Workshop on the Future Opportunities for Genome Sequencing and Beyond	Mark Gerstein Richard Myers
5:00 p.m.	**General Discussion**	
6:00 p.m.	**Working Dinner**	
7:00 p.m.	**Topic #1: Identifying and characterizing functional elements** Moderator: Carol Bult	
	7:00 p.m. – 7:30 p.m. The regulatory landscape: where are the gaps?	Ross Hardison
	7:30 p.m. – 7:45 p.m. Creating a framework for mechanistic studies	B. Franklin Pugh
	7:45 p.m. – 8:00 p.m. ENCODE 2.0: improving the syntax for understanding functional elements in the genome	Laurie Boyer
	8:00 p.m. – 8:15 p.m. Genomics at the "quantum" level: new directions for genomic data generation and functional validation	William Greenleaf
8:15 p.m.	**Topic #1 Discussion**	
9:00 p.m.	**Adjourn**	

Wednesday, March 11, 2015

8:00 a.m.	**Topic #2: Using genomic assays of function to interpret the role of genetic variation in disease** Moderator: Eric Boerwinkle	
	8:00 a.m. – 8:25 a.m. Leveraging whole genome annotation for genotype-phenotype association studies	Eric Boerwinkle

	8:25 a.m. – 8:50 a.m. Hirschsprung disease consequent to mutations in the RET gene regulatory network	Aravinda Chakravarti
	8:50 a.m. – 9:05 a.m. Genetically predicted endophenotypes: getting to the next level in understanding how genome variation drives disease	Nancy Cox
	9:05 a.m. – 9:20 a.m. Identification of regulatory variation important for maternal metabolism during pregnancy	William Lowe

9:20 a.m. **Topic #2 Discussion**

10:05 a.m. *Break*

10:30 a.m. **Topic #3: Using genomic assays of function to study basic biological questions**
Moderator: Aviv Regev

	10:30 a.m. – 10:50 a.m. Understanding basic biology using functional genomics: solving the genotype-phenotype problem	Brenda Andrews
	10:50 a.m. – 11:10 a.m. Fundamental insights into gene regulation from genomic analyses: past successes and future challenges	Karen Adelman
	11:10 a.m. – 11:30 a.m. Analyzing cytosine modifications in genomic DNA	Anjana Rao
	11:30 a.m. – 11:50 a.m. Ask not what you can do for ENCODE – ask what ENCODE can do for you	John O'Shea

11:50 a.m. **Topic #3 Discussion**

12:30 p.m. *Lunch*

1:30 p.m. **Discussion**
Moderators: Eric Boerwinkle, Carol Bult, Aviv Regev

2:30 p.m. **Final Recommendations, including priorities and balance of activities**
Moderators: NHGRI

3:30 p.m. **Adjourn**

National Human Genome Research Institute

From Genome Function to Biomedical Insight: ENCODE and Beyond

March 10-11, 2015

Natcher Conference Center
National Institutes of Health
Bethesda, Maryland

PARTICIPANT LIST

Karen Adelman, Ph.D.
Senior Investigator
National Institute of Environmental Health Sciences
National Institutes of Health
(919) 541-0001
adelmank@niehs.nih.gov

Brenda Andrews, Ph.D.
Professor and Director
The Donnelly Centre
University of Toronto
(416) 978-8562
brenda.andrews@utoronto.ca

Ewan Birney, Ph.D.
Associate Director
Head of CTTV, GTL
The European Bioinformatics Institute
+44 1223 492645
birney@ebi.ac.uk

Olivier Blondel, Ph.D.
Program Director
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
(301) 451-7334
blondelol@niddk.nih.gov

Eric Boerwinkle, Ph.D.
Professor and Chair
Human Genetics Center and Department of Epidemiology
University of Texas Health Science Center
Associate Director
Human Genome Sequencing Center
Baylor College of Medicine
(713) 500-9816
eric.boerwinkle@uth.tmc.edu

Laurie Boyer, Ph.D.
Sizer Career Development Associate Professor
Department of Biology
Massachusetts Institute of Technology
(617) 324-3335
lboyer@mit.edu

Lawrence Brody, Ph.D.
Senior Investigator
National Human Genome Research Institute
National Institutes of Health
(301) 496-7824
lbrody@mail.nih.gov

Lisa Brooks, Ph.D.
Program Officer
National Human Genome Research Institute
National Institutes of Health
(301) 435-5544
lisa.brooks@nih.gov

Carol Bult, Ph.D.
Professor and Deputy Director
Mammalian Genetics
The Jackson Laboratory Cancer Center
(207) 288-6324
carol.bult@jax.org

George Carlson, Ph.D.
Director and Professor
McLaughlin Research Institute
Montana State University
University of Montana, Benefits Health System
(406) 454-6044
gac@mri.montana.edu

Aravinda Chakravarti, Ph.D.
Director
Center for Complex Disease Genomics
Professor of Medicine, Pediatrics, Molecular Biology and Genetics and Biostatistics
Johns Hopkins University School of Medicine
(410) 502-7525
aravinda@jhmi.edu

James N. Coulombe, Ph.D.
Program Director, Developmental Genetics and Systems Developmental Biology
Developmental Biology and Structural Variation Branch
Eunice Kennedy Shriver National Institute of Child Health and Human Development
National Institutes of Health
(301) 451-1390
coulombej@mail.nih.gov

Julie Coursen
Program Analyst
Division of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
(301) 435-5662
julie.coursen@nih.gov

Nancy Cox, Ph.D.
Professor of Genetics and Medicine
Director
Vanderbilt Genetics Institute
Director
Division of Genetic Medicine
Vanderbilt University
(615) 322-2091
nancy.j.cox@vanderbilt.edu

Valentina Di Francesco, M.S.
Program Director, Computational Biology
National Human Genome Research Institute
National Institutes of Health
(301) 496-7531
vdifrancesco@mail.nih.gov

Joseph Ecker, Ph.D.
Professor
Howard Hughes Medical Institute
The Salk Institute
(858) 453-4100
ecker@salk.edu

Elise Feingold, Ph.D.
Program Director, Genome Analysis
Division of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
(301) 496-7531
elise_feingold@nih.gov

Adam Felsenfeld, Ph.D.
Program Director
National Human Genome Research Institute
National Institutes of Health
(301) 496-7531
adam_felsenfeld@nih.gov

Mark Gerstein, Ph.D.
Professor
Gerstein Laboratory
Yale University
(203) 432-6105
mark@gersteinlab.org

Daniel Gilchrist, Ph.D.
Program Director
National Human Genome Research Institute
National Institutes of Health
(301) 402-1966
daniel.gilchrist@nih.gov

Peter Good, Ph.D.
Program Director
Division of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
(301) 435-5796
goodp@mail.nih.gov

Eric Green, M.D., Ph.D.
Director
National Human Genome Research Institute
National Institutes of Health
(301) 496-0844
egreen@nhgri.nih.gov

William Greenleaf, Ph.D.
Assistant Professor, Genetics, Applied Physics
Stanford University
(650) 725-3672
wjg@stanford.edu

Mark Samuel Guyer, Ph.D.
Consultant
National Institutes of Health
(301) 435-5536
guyerm@exchange.nih.gov

Ross C. Hardison, Ph.D.
T. Ming Chu Professor of Biochemistry and Molecular Biology
Director
Center for Comparative Genomics and Bioinformatics
Department of Biochemistry and Molecular Biology
The Pennsylvania State University
(813) 863-0113
rch8@psu.edu

Carolyn M. Hutter, Ph.D.
Program Director
Division of Genomic Medicine
National Human Genome Research Institute
National Institutes of Health
(301) 451-4735
huttercm@mail.nih.gov

Donna Krasnewich, M.D., Ph.D.
Program Director
National Institute of General Medical Sciences
National Institutes of Health
(301) 594-0943
dkras@mail.nih.gov

Tuuli Lappalainen, Ph.D., M.S.
Assistant Professor
New York Genome Center and
Department of Systems Biology
Columbia University
(917) 753-2661
tlappalainen@nygenome.org

William Lowe, M.D.
Professor
Department of Medicine
Division of Endocrinology, Metabolism and Molecular Medicine
Feinberg School of Medicine
Northwestern University
(312) 503-2539
wlowe@northwestern.edu

Mathieu Lupien, Ph.D.
Scientist
University Health Network
Assistant Professor
University of Toronto
(416) 581-7434
mlupien@uhnres.utoronto.ca

Daniel MacArthur, Ph.D.
Assistant Professor
Massachusetts General Hospital
Broad Institute
Harvard Medical School
(617) 726-6028
macarthur@atgu.mgh.harvard.edu

Judy Mietz, Ph.D., M.S.
Program Director
Division of Cancer Biology
National Cancer Institute
National Institutes of Health
(240) 276-6250
mietzj@mail.nih.gov

Richard Myers, Ph.D.
Director and President
HudsonAlpha Institute for Biotechnology
(256) 327-0431
rmyers@hudsonalpha.org

Hannah Naughton
Scientific Program Analyst
Division of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
(301) 594-6563
hannah.naughton@nih.gov

John J. O'Shea, M.D.
Scientific Director
Chief
Lymphocyte Cell Biology Section
Molecular Immunology and Inflammation Branch
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institutes of Health
(301) 496-6026
osheajo@mail.nih.gov

Mike Pazin, Ph.D.
Program Director, Functional Genomics
Division of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
(301) 476-7531
pazinm@mail.nih.gov

Dana Pe'er, Ph.D.
Associate Professor
Departments of Biological Sciences and Systems Biology
Columbia University
(617) 372-0537
dpeer@biology.columbia.edu

Rudy O. Pozzatti, Ph.D.
Scientific Review Officer
Division of Extramural Operations
National Human Genome Research Institute
National Institutes of Health
(301) 402-8739
pozzattr@exchange.nih.gov

Frank Pugh, Ph.D.
Willaman Professor of Molecular Biology
Department of Biochemistry and Molecular Biology
Pennsylvania State University
(814) 863-8252
bfp2@psu.edu

Anjana Rao, Ph.D.
Professor and Division Head
La Jolla Institute for Allergy and Immunology
Sanford Consortium for Regenerative Medicine
University of California, San Diego
(858) 952-7161
arao@liai.org

Aviv Regev, Ph.D.
Principal Investigator
Professor and Investigator
Broad Institute
Massachusetts Institute of Technology and Howard Hughes Medical Institute
(617) 714-7021
aregev@broadinstitute.org

John Satterlee, Ph.D.
Program Director
National Institute on Drug Abuse
National Institutes of Health
(301) 435-1020
satterleej@nida.nih.gov

Jeffery A. Schloss, Ph.D.
Director
Division of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
(301) 496-7531
schlossj@exchange.nih.gov

Geetha Senthil, Ph.D.
Program Officer
National Institute of Mental Health
National Institutes of Health
(301) 402-0754
senthilgs@mail.nih.gov

Jay A. Shendure, M.D., Ph.D.
Professor
Genome Sciences
University of Washington
(206) 221-7377
shendure@uw.edu

Adam C. Siepel, Ph.D., M.S.
Professor and Chair
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
(516) 367-6922
asiepel@cshl.edu

Mike Snyder, Ph.D.
Department Chair and Professor
Department of Genetics
Stanford University
(650) 723-4668
mpsnyder@stanford.edu

John Stamatoyannopoulos, M.D.
Associate Professor of Genome Sciences and Medicine
Department of Genome Sciences
School of Medicine
University of Washington
(206) 685-2672
jstam@u.washington.edu

Paul Sternberg, Ph.D.
Thomas Hunt Morgan Professor of Biology
California Institute of Technology
(626) 395-2181
pws@caltech.edu

Hendrik Gerard Stunnenberg, Sc.D.
Professor and Head
Department of Molecular Biology
Scientific Director
Research Institute
Radboud University Nijmegen, The Netherlands
+31 654312535
h.stunnenberg@ncmls.ru.nl

Olga Troyanskaya, Ph.D.
Deputy Director for Genomics
Simons Foundation
Professor of Computer Science and Genomics
Lewis-Sigler Institute for Integrative Genomics
Princeton University
(646) 751-1284
ogt@genomics.princeton.edu

K-T Varley, Ph.D.
Assistant Professor
Huntsman Cancer Institute
University of Utah
(607) 592-0757
kt.varley@hci.utah.edu

Haoyi Wang, Ph.D.
Distinguished Visiting Professor
The Jackson Laboratory
(207) 288-6816
haoyi.wang@jax.org

ENCODE 3 Summary

March 6, 2015

The goal of the Encyclopedia of DNA Elements (ENCODE) project is to identify functional elements of the genome and thus providing a framework for interpreting the human genome in the context of human biology and disease. The ENCODE Consortium (Figure 1) consists of Data Production Centers responsible for efficiently generating specific data types, a Data Coordination Center (DCC) for data curation and public release, a Data Analysis Center (DAC) and an Analysis Working Group (AWG) to assist in and lead integrated analyses of the data, and computational analysis and technology development groups who have continuously developed and piloted new technologies and computational approaches. Together, the consortium has generated a product that is far greater than the sum of the parts. Key aspects of this are the development of high and uniform standards for data generation and data quality, uniform, cloud-based data processing pipelines that are transparent, reproducible, and available for use by users outside the consortium, the rapid and restriction-free release of verified datasets as soon as they are generated, and extensive educational and outreach efforts. Below we briefly highlight major accomplishments of the ENCODE project, followed by a more detailed overview of each activity of the ENCODE consortium.

Core accomplishments of the ENCODE project to date can be summarized as follows:

- Creation of vast, accessible catalogs of regulatory DNA, transcription factor occupancy and histone modification patterns, RNA binding protein occupancy, and RNA transcripts, as well as a standard curation of protein-coding and non-coding genes (GENCODE).
- Development and dissemination of standards and experimental methods for producing high-quality, reproducible data in a cost efficient manner from major assay types including ChIP-seq, RNA-seq, and DNase-seq.
- Development and dissemination of algorithms and software for analysis of major regulatory genomic data types, as well as tools and methods for integrating functional genomic data types.

Key biological insights provided by ENCODE:

- The biochemical features, structural and functional diversity, and evolution of regulatory DNA
- The systematic analysis and interpretation of non-coding disease- and trait-associated human variation, e.g., brought to light by GWAS studies
- The organization and architectural principles of transcriptional regulatory networks and circuits
- Relationships between regulatory DNA and higher-order chromatin features and interactions
- The organization, diversity, and nuclear compartmentalization of RNA transcription and its interplay with chromatin and regulatory DNA.

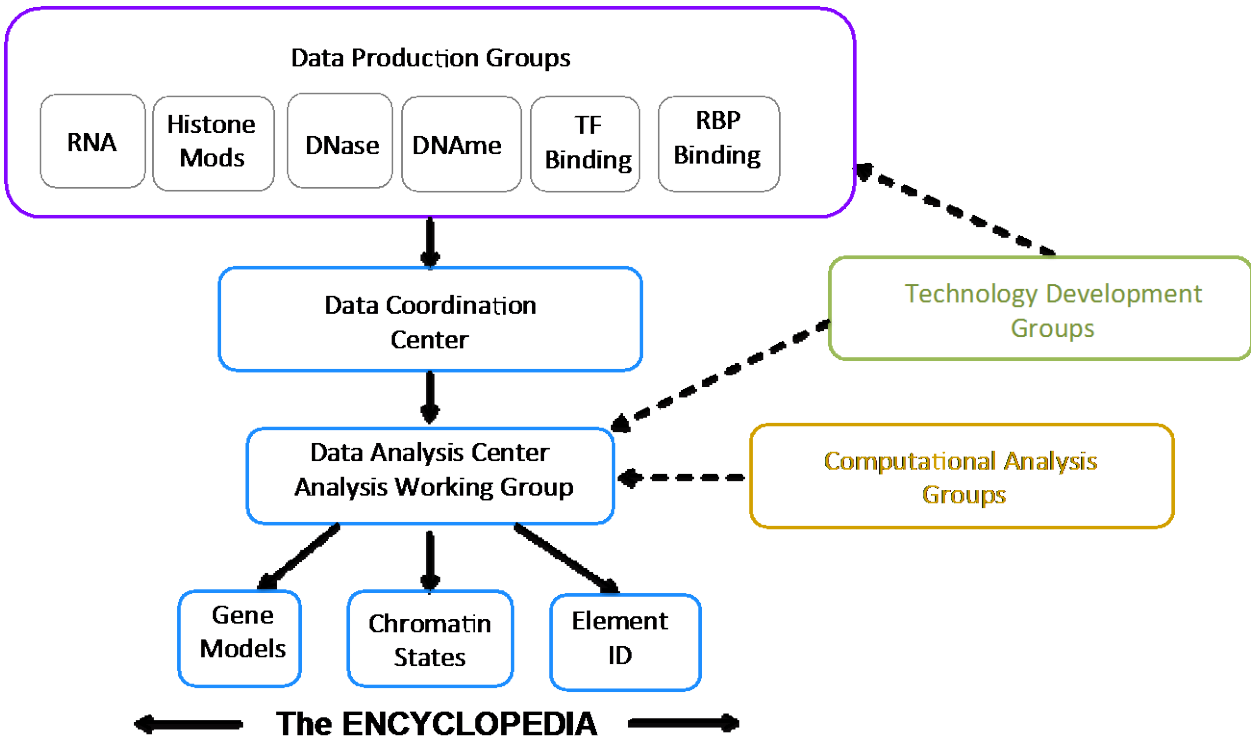

Figure 1. Overview of the organization of the ENCODE consortium.

General Impact

Just as the impact of the Human Genome Project took several years to be realized, the long-term impact of ENCODE has yet to fully mature. There are, however, many components of the project suggesting that ENCODE will have a long-lasting legacy. For example, the ENCODE consortium has a strong track record of publishing high impact papers that have been cited thousands of times each (e.g., the main consortium paper published in *Nature* in 2012 has already been cited 2,379 times according to Google Scholar). To date, ENCODE has published 520 papers, and modENCODE and mouseENCODE have published 162 and 26, respectively. More importantly, however, is the use of ENCODE data beyond the immediate members of the ENCODE consortium. To date, about 750 papers by authors without ENCODE funding have used ENCODE data in their studies of human disease, basic biology, or methods development. Key to the uptake of ENCODE data beyond the consortium is the fact that the datasets are rapidly released to the public, of high quality, are generated and processed in a consistent manner, and have few if any restrictions on community use. Community use should also increase now that the ENCODE data processing pipelines are shared through GitHub and the DNA Nexus cloud. Furthermore, ENCODE is closely working with related projects to increase standardization of data processing, metadata, and APIs across projects. This will facilitate integrated analysis of data between projects further enhancing impact. Accessibility to ENCODE data is easier than ever and available on the new ENCODE portal which allows users to search, download, and display raw data and processed data. To further enhance uptake, ENCODE provides tutorials on use of the resource and will offer a users meeting in the summer of 2015. The ENCODE portal also shares the data standards documents to enhance reproducibility and

to allow others to adopt them if they wish. ENCODE has collaborated with many groups including the GENEVA project, CHARGE and eMERGE. A number of projects have been inspired by ENCODE; a zebrafish project (ZENCODE), an agricultural animals project (FAANG), and a psychiatric diseases project (PsychENCODE). Finally, in addition to the individual datasets, technologies, software tools, and publications, the main “product” of ENCODE will be the Encyclopedia of DNA Elements – an annotation of functional elements of the genome. The Encyclopedia will provide the framework for interpreting the human genome and how variation impacts human biology and disease.

Data Production

ENCODE has generated 4,294 genome-wide experiments (nearly all of which consist of two replicates) that have been publicly released or submitted to the DCC for imminent public release. ENCODE projects to complete an additional 6,221 replicated experiments by the end of the project for a total of 10,515, approximately 84% of which are human with the remainder from mouse.

ENCODE has employed a diverse collection of assays to interrogate protein-DNA interactions (TF ChIP-Seq), chromatin structure (histone modification ChIP, DNase-Seq and ATAC-Seq), DNA methylation (whole-genome shotgun bisulfite sequencing, RRBS, and arrays), protein-RNA interactions (iCLIP and RIP-Seq), and RNA transcripts (RNA-Seq for long, short and small RNAs, poly(A)+ and total) to name a few. Collectively, these assays have been used to profile over 500 distinct biological samples, 76% of which are human. Most samples analyzed from mouse are either tissues or primary cells. While presently about half of the completed human samples are from immortalized cell lines, it is projected that by the end of ENCODE3, most human samples will be from tissues or primary cells, in part through a collaboration between ENCODE and GTEx. Over 200 assays have also been conducted in stem cells or induced pluripotent stem cells and over 150 assays in differentiated stem cells or induced pluripotent stem cells. A small number of assays have been conducted in cells treated with various perturbagens (e.g., estradiol, ethanol, Sendai virus, tumor necrosis factor, etc.) though this is a largely unexplored area for ENCODE.

Among all biosamples, a small number have been deeply sampled with many assays, while a large number have been profiled with a small number of assays. ENCODE is an associate member of the International Human Epigenome Consortium (IHEC), and we project we will complete about 25 human and 50 mouse complete IHEC epigenomes (comprised of at least DNAme WGBS, mRNA-seq, and ChIP-Seq for H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K27ac, and H3K9me3 data). By the end of ENCODE3, it is projected that K562 and HepG2 cells will have been subjected to 1,745 and 1,495 assays, respectively. These assays include complete epigenomes, ChIP-Seq and iCLIP assays for hundreds of transcription factors and RNA binding proteins with DNA and RNA, knockdown experiments followed by RNA-Seq to assess the functional impact of the observed protein-DNA and protein-RNA interactions, as well as other assays providing information on chromatin accessibility, DNA replication timing and 3D chromatin interactions. Together, these incredibly deep and uniformly

generated datasets will provide an unprecedented opportunity to model the regulatory networks in these cells. At the other end of the spectrum, approximately 20% of all biosamples have been subjected to only a single genome-wide assay, the majority of which are either RNA-Seq or DNase-Seq, some of which are very widely used by the community.

By the end of ENCODE3, it is projected that over 1,500 different human proteins (or modified proteins) and 69 mouse proteins, will have been assayed by at least one protein-nucleic interaction assay (ChIP-Seq, iCLIP, etc.) in at least one biological sample. Most experiments targeting individual proteins (28%) have focused on major histone modification marks (H3K4me3, H3K27me3, H3K36me3, H3K4me1, H3K27ac, H3K9me3, etc.) as well as the architectural TF CTCF, the large subunit of RNA polymerase II and the histone acetyltransferase p300. The majority of the remaining targets are either transcription factors or RNA binding proteins which have been studied in only a small number of biological samples. Although ENCODE has studied an extremely large number of proteins, hundreds of DNA and RNA binding proteins have not yet been studied.

An extremely important contribution of ENCODE is the generation, identification, and characterization of high-quality reagents used in the project. Using quality criteria that have set a standard in the field, over 500 human and 60 mouse antibodies have been validated for use in immunoprecipitation experiments and it is projected that several hundred additional antibodies will be validated by the end of ENCODE3. In addition, many cell lines expressing epitope tagged proteins have been generated and hundreds of shRNAs that efficiently deplete the target mRNAs have been validated.

ENCODE3 has begun to systematically validate the function of elements identified from the vast array of genome-wide datasets that have been generated. Specifically, several groups are predicting enhancer elements and testing their activity in transgenic mouse enhancer assays and high throughput reporter assays. Such efforts are critical to differentiate elements that have biochemical activity (e.g., can be identified in genome-wide assays) versus those that are biologically functional.

Data Coordination Center (DCC)

During ENCODE3, the Data Coordination Center (DCC), working with production labs, DAC and the AWG, defined a new metadata standard for describing high-throughput sequencing assays and computational analyses, engineered robust uniform processing pipelines that processed these data, and built a new portal for the ENCODE Consortium that allows the scientific community to access these metadata and data as well as serves as a hub for sharing and communication among Consortium members. The implementation of these 3 major deliverables at the ENCODE DCC focused on maintaining interoperability with other genomic resources, capturing data provenance to ensure reproducibility, and providing novel ways to access ENCODE metadata and data.

The new metadata standard defines a data model that can handle 40+ high-throughput assay types in fly, worm, mouse, and human, by capturing key experimental variables, such as biosamples and assay methods. Interoperability between genomic resources is maintained through the use of ontologies also employed by EBI ArrayExpress to capture the experimental variables. In addition, the DCC provides outreach to other projects such as the Nuclear Receptor Signaling Atlas (NURSA), Reference Epigenomics Mapping Consortium (REMC), International Human Epigenome Consortium (IHEC), and GTEx, among others, to promote the use of ontologies for capturing experimental variables for increased interoperability. The metadata standard also promotes data provenance and reproducibility in that the data model supports linking biological samples that originate from a single donor as well as supports the description of software used for a pipeline, parameters used in running that software, and the identification of files that were used as input or output. To provide novel methods to access ENCODE metadata, it is available as a data model in JSON-LD which supports semantic web queries.

ENCODE3 has engineered and distributed uniform processing pipelines in order to promote data provenance & reproducibility as well as allow interoperability between genomic resources. The pipelines have defined metadata that promotes data provenance & reproducibility. All data files, reference genome versions, software versions, and parameters used by pipeline are captured and available via the ENCODE Portal. Pipelines are publicly available so a diverse range of biomedical researchers have access to and can run the exact pipeline that are used to generate ENCODE results. ENCODE pipelines maintained and used by the DCC are freely available for public use at DNAnexus.com via web browser, or linux command line via the ENCODE DCC github, so that researchers can process their data with “the” ENCODE pipelines, or create modified versions of the pipelines to suit their needs. All ENCODE primary and processed data are available and distributed without charge via the Amazon Web Services (AWS). This allows traditional download but also provides access to the complete data warehouse from an account at AWS. Access to the pipelines and data via the cloud allows even small labs the ability to use the data or software without access to institutional compute clusters. Additionally, the DCC is providing outreach to IHEC, California Institute for Regenerative Medicine (CIRM), and bioinformatics core groups to encourage use of common processing pipelines.

A new ENCODE Portal (encodeproject.org) has been created to promote use of ENCODE data and results by the scientific community and serve as a hub for sharing and communication among Consortium members. The portal pages list the antibody characterization and data standards defined by the Consortium, providing transparency about the methods and standards supporting the assays and analysis performed by the Consortium. The portal also displays the relationship between donors and biosamples as well as graphical displays of pipelines, software used for pipelines, and files generated by the pipeline encourage data reuse. In addition, the Portal provides an integration point for other significant consortium data, for example the REMC and the Genomics of Gene Regulation (GGR) results and metadata integrated with ENCODE products. Development of the portal has included novel ways to access ENCODE data. The ENCODE Portal is based on a REST API which is an industry

standard for interacting with websites and other databases. The development of an ENCODE REST API provides programmatic data submission by members of the Consortium as well as programmatic data retrieval for the scientific community. The portal provides innovative search features utilizing the structured metadata as well as the ability to search processed ENCODE data by a genomic coordinate (eg, a variant), a genomic region (eg, a coordinate range), or a gene name plus flanking region (eg, a gene name +/- 10 kb) in order to retrieve data files that contain ENCODE-defined elements in that region. The ENCODE Encyclopedia of Elements will also be integrated within this view of the data. ENCODE assays can be searched using faceted browsing to narrow down the list of assays interesting to the user. The Portal also includes alternatives to visualize ENCODE data, including visualization of files from any search or arbitrary set of files at the UCSC Genome Browser via a track hub and graphical summaries of the ENCODE encyclopedia from search results. This allows peak files from only a subset of ChIP-seq assays to be displayed instead of the whole set of transcription factor ChIP-seq assays. To promote usage of the data, results, pipelines and files are stored in the cloud. This provides transparent and open access to ENCODE results, achievements and deliverables allowing the data to be reused, analyses queried and integrated, software to be reused, and pipelines to be effortlessly applied to biomedical research.

Data Analysis Center (DAC) and Analysis Working Group (AWG)

ENCODE 3 maintains a Data Analysis Center (DAC), as a component of the ENCODE Data Coordination and Analysis Center (EDCAC). DAC members are tasked with coordinating and assisting in the integrative analysis of data produced by the ENCODE Consortium, developing pipelines for data processing, and working with the DCC component of the EDCAC to integrate them for automatic application on all datasets. In addition, DAC members have developed independent software tools for general use (Appendix 1). The DAC is responsible for coordinating the analyses required to generate the Encyclopedia of DNA elements - the main product of ENCODE.

The Analysis Working Group (AWG) is composed of individuals from several of the data production and analysis projects and was responsible for the integrated analysis of multiple datatypes. The AWG led many of the analyses that resulted in the high impact consortium papers.

Computational Analysis Groups

One major difference between ENCODE3 and the previous phases of the project was the inclusion of six groups who are funded to generate innovative computational tools and approaches to analyze ENCODE data. These projects have developed new statistical and computational approaches to reduce the complexity of ENCODE data, allow comparisons involving many ENCODE datasets at once, identify regulatory elements in the human genome, including in repetitive elements, to determine how regulatory elements work together, integrate ENCODE data with GWAS data, determine how changes in DNA sequence lead to changes in gene expression, and identify genetic differences that alter RNA processing. To date, the computational analysis groups have generated at least 25 software tools that are currently being

used to analyze ENCODE data. The software tools used and developed by the ENCODE Consortium are listed in the section "Analysis tools generated by ENCODE" and on the ENCODE portal (https://www.encodeproject.org/software).

Technology Development Groups

In 2012, 11 groups were funded to develop revolutionary technologies to help identify genomic elements that play a role in determining what genes are expressed and at what levels in different cells. These technology development areas were focused in three areas – the discovery of functional elements, the characterization of functional elements, and computational analyses. Together these projects led to the development of new technologies that facilitated the identification of branchpoints involved in pre-mRNA splicing, measuring mRNA degradation and splicing kinetics, improving the power of ChIP-Seq, high throughput assays to validate a variety of functional elements, and computational approaches to model cell-specific gene expression programs and chromatin structure. The technologies developed in these projects have not only had an important impact on ENCODE, but have also been broadly adapted by researchers outside the consortium. The ENCODE3 technology development groups have published 27 papers, and 141 throughout the entire project.

Analysis tools generated by ENCODE

ACT: The aggregation and correlation toolbox (ACT) is an efficient, multifaceted toolbox for analyzing continuous signal and discrete region tracks from high-throughput genomic experiments, such as RNA-seq or ChIP-chip signal profiles from the ENCODE and modENCODE projects, or lists of single nucleotide polymorphisms from the 1000 genomes project. It is able to generate aggregate profiles of a given track around a set of specified anchor points, such as transcription start sites. It is also able to correlate related tracks and analyze them for saturation--i.e. how much of a certain feature is covered with each new succeeding experiment. The ACT site contains downloadable code in a variety of formats, interactive web servers (for use on small quantities of data), example datasets, documentation and a gallery of outputs. http://act.gersteinlab.org/ PMID: 21349863

AlleleSeq: A computational pipeline that is used to study allele-specific expression (ASE) and allele specific binding (ASB). The pipeline first constructs a diploid personal genome sequence, then maps RNA-seq and ChIP-seq functional genomic data onto this personal genome. Consequently, locations in which there are differences in the number of mapped reads between maternally- and paternally-derived sequences can be identified, thereby providing evidence for allele-specific events. http://alleleseq.gersteinlab.org/home.html PMID: 21811232

ASARP: Allele-Specific Alternative mRNA Processing; This is a method to identify SNPs that influence mRNA processing. An updated version of this software is released; https://github.com/cyruschan/ASARP PMID: 22467206 PMCID: PMC3401465

atSNP: a fast importance sampling method for evaluating binding affinity changing potential of SNPs. https://github.com/chandlerzuo/atSNP Manuscript in preparation.

BETA: integrating ChIP-seq binding and differential expression (i) to predict whether the factor has activating or repressive function; (ii) to infer the factor's target genes; and (iii) to identify the motif of the factor and its collaborators, which might modulate the factor's activating or repressive function. http://cistrome.org/BETA/ PMID: 24263090

CCM: Cooperative Chromatin Model (CCM) - predicts chromatin accessibility (DNase-seq data) from novel DNA sequence. Evaluates every base of the genome for its estimated importance for controlling accessibility (manuscript under review)

Census: Tool to estimate sequencing library complexity from test sequencing runs (released, manuscript in preparation)

cnvCSEM: Copy number variation (CNV) guided multi-read allocation. Software available from http://www.stat.wisc.edu/~qizhang/ PMCID: PMC4184254

curveHist: a functional curve testing approach for identifying differential histone modifications. Manuscript and software in preparation.

dPeak: a model-based tool for identifying closely located binding events from ChIP-seq and ChIP-exo data.https://github.com/dongjunchung/dpeak PMCID: PMC3798280

Fit-Hi-C: Assigning statistical confidence estimates to Hi-C data. PMCID: PMC4032863

FixSeq: Method to adjust read counts so they are not overdispersed. An alternative to simple de-duplication of reads. Improves the performance of many high-throughput analysis packages (released, https://bitbucket.org/thashim/fixseq) PMCID: PMC3945112

FusionSeq: FusionSeq may be used to identify fusion transcripts from paired-end RNA-sequencing. FusionSeq includes filters to remove spurious candidate fusions with artifacts, such as misalignment or random pairing of transcript fragments, and it ranks candidates according to several statistics. It also includes a module to identify exact sequences at breakpoint junctions. http://archive.gersteinlab.org/proj/rnaseq/fusionseq/ PMID: 20964841

GCAP: Data QC and analysis pipeline for DNase-seq analysis: https://github.com/qingian/GCAP

GEM: High resolution ChIP seq event caller. (released, in use by DCC) http://sysbio.mit.edu/gem/ High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. Yuchun Guo, Shaun Mahony & David K Gifford, (2012) PLoS Computational Biology 8(8): e1002638. DOI: 10.1371/journal.pcbi.1002638 PMCID: PMC3415389

GIREMI: Genome-independent Identification of RNA Editing by Mutual Information; This is a new method to predict RNA editing sites using RNA-Seq data alone without the need of genome sequencing data. doi:10.1038/nmeth.3314 PMID: 25730491

GoShifter: Software for testing if a set of SNPs are enriched in particular functional annotations of the genome through peak shifting the functional annotations. https://www.broadinstitute.org/mpg/goshifter/ Preprint at http://biorxiv.org/content/early/2014/09/18/009258

GRIT: (http://grit-bio.org/), a tool for the integrative analysis of RNA-seq, CAGE, PAS-seq, RAMPAGE, and other RNA datatypes. PMID: 24633242.

GSC: Genome Structure Correction, (http://projecteuclid.org/euclid.aoas/1294167794), a tool for assessing when two or more features defined on genomes are more associated than expected by chance alone.

HAYSTACK: chromatin state variation and cell-type specific regulators. PMCID: PMC3903219

IDR: irreproducible Discovery Rate, (https://www.encodeproject.org/software/idr/), a tool for the assessment of the reproducibility of results in high-throughput studies, analygous to the False Discovery Rate.

iGRAS: intronic tag SNPs for Genetic Regulation of Alternative Splicing; This is a method to identify intronic SNPs that are involved in causing splicing alteration using RNA-Seq data obtained from cellular fractions (nucleus vs. cytosol). (manuscript under review)

ILM: Introspective Learning Machines, a tool for feature detection, classification, and regression. Includes Unconstrained Surface Mapping (**USM**). In production; publication pending.

IQseq: A tool for isoform quantification with RNA-seq data. Given isoform annotation and alignment of RNA-seq reads, it will use an EM algorithm to infer the most probable expression level for each isoform of a gene. http://archive.gersteinlab.org/proj/rnaseq/IQSeq/ PMID: 22238592

jMOSAiCS: Joint peak calling and inferring patterns of binding/modification across many ChIP-seq datasets. Zeng X (2013). jmosaics: Joint analysis of multiple ChIP-Seq data sets. R package version 1.6.0. http://www.bioconductor.org/packages/release/bioc/html/jmosaics.html

KMAC: De novo motif discovery method - Discover enriched motifs (in KSM and PWM representation) from a set of sequences (typically TF ChIP-seq data), or enriched in a set of sequences comparing to another set of sequences. (manuscript in preparation)

Loregic: a computational method integrating gene expression and regulatory network data to characterize the logical cooperativity of regulatory factors. Loregic uses all 16 possible two-input-one-output logic gates (e.g. AND or XOR) to describe triplets of two factors regulating a common target, and finds the gate that best matches each triplet's observed gene expression pattern across many conditions. Using human ENCODE ChIP-Seq and TCGA RNA-Seq data, we are able to demonstrate how Loregic characterizes complex circuits involving both proximally and distally regulating transcription factors (TFs) and also miRNAs. github.com/gersteinlab/loregic PMID: *(in press)*

Mango: A complete ChIA-PET data analysis pipeline that provides statistical confidence estimates for interactions and corrects for major sources of bias including differential peak enrichment and genomic proximity. (Manuscript under review)

MBASIC: a generative model for analyzing and grouping multiple loci based on ENCODE ChIP-seq data. These loci can be a collection of motif locations, a liberal set of SNPs, or peaks from an ENCODE or non-ENCODE experiment. (Manuscript under review)
Software: https://github.com/chandlerzuo/mbasic

MOSAiCS-HMM: Peak caller specific for Histone ChIP-seq data (builds on and extends our peak caller MOSAiCS). Software: https://github.com/dongjunchung/mosaics

MultiGPS: Multi-condition ChIP seq event caller that aligns events across conditions and permitting the detection of differential events. http://mahonylab.org/software/multigps/ PMCID: PMC3967921

MUSIC: An algorithm for identification of enriched regions at multiple scales in the read depth signals from ChIP-Seq experiments. MUSIC first filters the ChIP-Seq read-depth signal for systematic noise from non-uniform mappability, which fragments enriched regions. It then performs a multiscale decomposition, using median filtering, identifying enriched regions at multiple length scales. https://github.com/gersteinlab/MUSIC PMID: 25292436

Pastis: Inferring 3D structure from Hi-C data. PMCID: PMC4229903

PeakSeq: A tool for calling peaks corresponding to transcription factor binding sites from ChIP-Seq data scored against a matched control such as input DNA. PeakSeq employs a two-pass strategy in which putative binding sites are first identified in order to compensate for genomic variation in the 'mappability' of sequences, before a second pass filters out sites not significantly enriched relative to the normalized control, computing precise enrichments and significances. http://info.gersteinlab.org/PeakSeq PMID: 19122651

Perm-seq: a probabilistic read mapping method that can supervise multi-read allocation in TF ChIP and related experiments.
Software: http://www.stat.wisc.edu/~keles/Software/perm-seq/ (Manuscript under review)

PIQ: Method to resolve protein-DNA binding from DNase-seq data. Can produce binding calls for hundreds of different factors from a single DNase-seq experiment http://piq.csail.mit.edu (released)

pRSEM: Prior-enhanced version of RNA-seq quantification method RSEM (RSEM is in use by DCC). Manuscript and software in preparation.

RABIT: Regression analysis with background integration. Fast feature selection and regression method integrating ENCODE ChIP-seq, TF motifs, RBP motifs and TCGA expression, CNV, and DNA methylome data to identify key TFs and RBPs that regulate gene expression changes in different tumors. Paper under revision at PNAS, website http://rabit.dfci.harvard.edu will be available by ENCODE 2015 consortium meeting date.

RASER: Reads Aligner for SNPs and Editing sites of RNA; This is a new read aligner customized for accurate mapping of reads harboring SNPs or RNA editing sites. (manuscript in preparation)

RSEQtools: A suite of tools that use Mapped Read Format (MRF) for the analysis of RNA-Seq experiments. MRF is a compact data format that enables anonymization of confidential sequence information while maintaining the ability to conduct subsequent functional genomics studies. RSEQtools provides a suite of modules that convert to/from MRF data and perform common tasks such as calculating gene expression values, generating signal tracks of mapped reads, and segmenting that signal into actively transcribed regions. http://archive.gersteinlab.org/proj/rnaseq/rseqtools/ PMID: 21134889

Segway: Performing semi-automated genome annotation on the basis of heterogeneous collections of genomic data, including histone modification, DNase sensitivity, TF binding, RNA expression, Hi-C, etc. PMCID: PMC3340533

SeqGL: Software to learn multiple sequence signals from ChIP-, DNase-, and ATAC-seq data. Manuscript under review. https://bitbucket.org/leslielab/seqgl/wiki/Home

spliceVAR: Method to identify regulatory networks (SNPs and proteins) that underlie the variation of alternative splicing events. (manuscript in preparation)

Sprout: ChIA-PET interaction analysis tool with improved detection capability (released)

Statmap: (http://statmap-bio.org/) a tool for aligning short reads to repetitive genomes. PMID: 21177961.

Sushi: An R/Bioconductor package that allows flexible integration of genomic visualizations into highly customizable, publication-ready, multi-panel figures from common genomic data formats including Browser Extensible Data (BED), bedGraph and Browser Extensible Data Paired-End (BEDPE). http://bioconductor.org/packages/release/bioc/html/Sushi.html PMID: 24903420

UES (Uncovering Enrichment through Simulation): Software for testing if a set of SNPs are enriched in particular functional annotations of the genome through selection of random sets of SNPs. Manuscript in preparation.

WASP: WASP is a software package for two related tasks: (1) correcting allelic bias in mapped sequencing reads and, (2) identifying molecular quantitative trait loci (QTLs) using next-generation sequencing data (e.g. gene expression QTLs or histone mark QTLs). The WASP mapper works with any read mapping pipeline that outputs BAM or SAM format. WASP identifies molecular QTLs using a statistical test that combines information about the total depth and allelic imbalance of mapped reads. WASP can call QTLs with very small sample sizes (as few as 10) compared to traditional QTL mapping approaches. https://www.encodeproject.org/software/wasp/ van de Geijn B, McVicker G, Gilad Y, Pritchard J.. WASP: allele-specific software for robust discovery of molecular quantitative trait loci bioRxiv. 2014 Nov 7; doi:10.1101/011221

Projects With Similarities To ENCODE

Project Title	Project Acronym	URL	Funding Agencies	Sample Types	Assays/Data Types	Project Summary	Similarities with ENCODE	Distinctions from ENCODE	Status	Coordination with ENCODE	Program Contacts
International Human Epigenome Consortium	IHEC	http://www.ihec-epigenomes.org/	Consortium of projects funded by member nations	Healthy and diseased; generally purified cells	Transcriptomic; epigenomic	Data collection and reference maps of human epigenomes for key cellular states relevant to health and diseases	Chromatin and gene expression profiling of numerous samples; provides resources for interpretation of variant data	Disease samples	Active	ENCODE is IHEC associate member	Eric Marcotte (lead); Mike Pazin
Reference Epigenome Mapping Centers	REMC	http://www.roadmapepigenomics.org/	NIH Common Fund	Healthy tissues	Transcriptomic; epigenomic	Data collection, integrative analysis and a resource of human epigenomic data	Chromatin and gene expression profiling of numerous samples; provides resources for interpretation of variant data	Human only; producing reference epigenomes; limited number of assay types	Completed 2008-2013	REMC metadata hosted at ENCODE DCC; both IHEC members	Fred Tyson, Lisa Chadwick, John Satterlee (leads); Elise Feingold, Mike Pazin, Dan Gilchrist
BLUEPRINT	BLUEPRINT	http://www.blueprint-epigenome.eu/	EU	Normal and malignant haematopoietic cells	Transcriptomic; epigenomic	Data collection and analysis of normal and malignant blood cells	Chromatin and gene expression profiling of numerous samples; provides resources for interpretation of variant data	Disease samples; limited to haematopoietic cells; limited number of assay types	Active 2011-2016	Through IHEC membership	Henk Stunnenberg (lead); Mike Pazin
PsychENCODE	PsychENCODE	http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-14-020.html	NIMH	Normal and diseased neural tissues and cell types from human brain, and iPSC and CNON-derived neurons	Transcriptomic; epigenomic; genomic; proteomic	Data collection and integrative analysis of human neural epigenomic, genomic, transcriptomic and proteomic data	Cataloguing of functional elements; chromatin and gene expression profiling of numerous samples; provides resources for interpretation of variant data	Emphasis on normal and diseased neural tissues and cell types from adult and developing human brain, and comparison to non-human primates	Active 2014-present		Geetha Senthil
Functional Annotation of the Mammalian Genome	FANTOM	http://fantom.gsc.riken.jp	RIKEN	Human and mouse primary cells, tissues and cell lines	Transcriptomic; CAGE	Data collection of CAGE transcriptomic data and data analysis to annotate human and mouse functional elements	Gene expression profiling of numerous samples; focus on functional element annotation; provides resources for interpretation of variant data	No chromatin data	Active 2000-present	Includes four cell lines used in ENCODE	Yoshihide Hayashizaki, Piero Carninci
4D Nucleome	4DN	http://commonfund.nih.gov/4Dnucleome/index	NIH Common Fund	TBD	Imaging; genomic	Nuclear architecture; technology development and mapping projects; data integration and coordination center	Global interactions between gene loci and regulatory elements	Emphasis on tech dev; Imaging; nuclear dynamics; modeling of structure/function relationships; investigation of poorly characterized nuclear features	Beginning 2015	Program Officer Overlap	Olivier Blondell, Judy Meitz (leads); Mike Pazin

Projects With Similarities To ENCODE

Project Title	Project Acronym	URL	Funding Agencies	Sample Types	Assays/Data Types	Project Summary	Similarities with ENCODE	Distinctions from ENCODE	Status	Coordination with ENCODE	Program Contacts
Genomics of Gene Regulation	GGR	http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-13-012.html	NHGRI	Skin, immune system, nuclear hormone response	Transcriptomic; epigenomic	Determine how to develop predictive gene regulatory network models from genomic data	Chromatin and gene expression profiling of numerous samples	Goal is network modeling, based on data from cell transitions	Active 2015-2018	GGR data and metadata hosted at ENCODE DCC; Program Officer Overlap	Mike Pazin (lead)
Genotype-Tissue Expression Project	GTEx	http://www.gtexportal.org/home/	NIH Common Fund	post-mortem, non-diseased human tissues	WGS and WES; transcriptomic; additional data types in eGTEx (proteomic, epigenomic, ASE)	Data collection and analysis of variation in human gene expression, across individuals, and across >30 tissues from the same subjects	Provides resources for interpretation of variant data; gene expression profiling of numerous samples	Large population data-set; no chromatin data; not focused on identification of functional elements	Active 2010-present	Collaboration to perform ENCODE assays on GTEx samples	Simona Volpi, Jeff Struewing
Library of Integrated Network-based Cellular Signatures	LINCS	https://commonfund.nih.gov/LINCS/	NIH Common Fund	Primary cells; cell lines; iPS cells, differentiated neurons and cardiomyocytes	Transcriptomic; phosphoproteomic; imaging; epigenomic	Data collection and analysis of molecular signatures describing how different cell types respond to perturbing agents	Gene expression profiling of numerous samples	Emphasis on cataloguing responses to perturbations (small molecule, genetic, disease); not focused on identification of functional elements	Phase 1 2010-2013; Phase 2 2014-present		Ajay Pillai
International Cancer Genome Consortium	ICGC	http://www.icgc.org/	Consortium of projects funded by member nations	Tumor and normal	WGS and WES; transcriptomic; epigenomic	Data collection and analysis of genomic, transcriptomic and epigenomic changes in 50 different tumor types (includes TCGA samples)	Chromatin and gene expression profiling of numerous samples; provides resources for interpretation of somatic mutation data	Emphasis on cataloguing somatic genomic abnormalities; limited to tumors; not focused on identification of functional elements	Active 2008-present		Carolyn Hutter, Heidi Sofia
The Cancer Genome Atlas	TCGA	http://cancergenome.nih.gov/	NHGRI, NCI	Tumor and normal	WGS and WES; transcriptomic; epigenomic; proteomic	Data collection and analysis of genomic, transcriptomic, and epigenomic changes in ~30 different tumor types, and repository for DNA and RNA sequence data	Chromatin and gene expression profiling of numerous samples; provides resources for interpretation of somatic mutation data	Human only; tumors only; no chromatin data beyond DNAme; not focused on identification of functional elements	Active 2005-2016	Through ENCODE AWG Cancer working group	Carolyn Hutter, Heidi Sofia
Interpreting Variation in Human Non-Coding Genomic Regions Using Computational Approaches and Experimental Assessment	FunVar (will be updated)	http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-13-013.html	NHGRI, NCI, NIDA	Various	TBD; functional tests of prediction specificity/sensitivity	Development of computational approaches to interpret sequence variation in non-coding regions, and assessment of approaches through targeted data collection	Develops approaches for interpreting variant data	Goal is developing computational approaches to infer causal variants; limited data generation	Active 2015-2019	Program Officer Overlap	Lisa Brooks (Lead), Mike Pazin, Stefanie Nelson (NCI), Joni Rutter (NIDA)
Knockout Mouse Phenotyping Program	KOMP2	https://commonfund.nih.gov/KOMP2/	NIH Common Fund	Mice	Phenotypic	Data collection for standardized phenotyping of a genome-wide collection of mouse knockouts; member of International Mouse Phenotyping Consortium (IMPC)			Active 2011-2021		Colin Fletcher

ENCODE 2020: From Elements to Function

March 6, 2015

The ENCODE project has provided a framework for interpreting the human and mouse genomes. Beginning with the Pilot Project in 2003, ENCODE has taken a leading role in developing and implementing at scale technologies and platforms for delineating the genomic sequence segments that display the biochemical signatures of functional elements. From the initiation of the ENCODE production phase in 2007 to the present, the project has created an accessible, widely-used resource that is impacting our understanding of human genome function, and its connection with diverse aspects of human biology and disease. The development of this resource was critically enabled by a consortium model, which integrated data production centers focused on specific data types, a dedicated center for data curation and public release, and the continuous development and piloting of new technologies and computational approaches through dozens of R-series grants. ENCODE has also provided a model for other large-scale functional genomics efforts including the Roadmap Epigenomics Project, the International Human Epigenome Consortium, and others.

Core accomplishments of the ENCODE project to date can be summarized as follows:

- Creation of vast, accessible catalogs of regulatory DNA, transcription factor occupancy and histone modification patterns, and RNA transcripts, as well as a standard curation of protein-coding and non-coding genes (GENCODE).
- Development and dissemination of standards and experimental methods for producing high-quality, reproducible data in a cost efficient manner from major assay types including ChIP-seq, RNA-seq, and DNase-seq.
- Development and dissemination of algorithms and software for analysis of major regulatory genomic data types, as well as tools and methods for integrating functional genomic data.
- ENCODE has trained a new generation of fellows and students in genome science, who continue to play major roles in methods development, data generation and analysis.

In addition, Consortium members, frequently in collaboration with investigators from the broader community, have published pioneering applications and analyses of ENCODE data across a spectrum of biological paradigms. These have resulted in many foundational insights, and have catalyzed research in diverse areas including:

- The biochemical features, structural and functional diversity, and evolution of regulatory DNA
- The systematic analysis and interpretation of non-coding disease- and trait-associated human variation, e.g., brought to light by GWAS studies
- The organization and architectural principles of transcriptional regulatory networks and circuits
- Relationships between regulatory DNA and higher-order chromatin features and interactions
- The organization, diversity, and nuclear compartmentalization of RNA transcription and its interplay with chromatin and regulatory DNA.

ENCODE data have been extensively utilized by the scientific and biomedical community, as evidenced by thousands of publications using or citing ENCODE data and analysis.

Functional genomics: Imminent challenges and the role of ENCODE

Our understanding of the living human genome and its role in biology and disease is progressing rapidly but is still nascent. Despite rapid progress across the field of functional genomics, identifying all functional elements of the human genome is an unfulfilled aspiration. In fact, ENCODE data reveal greater diversity (combinatorial activation patterns and modification signatures) and greater numbers (up to millions) of elements than anticipated. Furthermore, we now appreciate that the relationships between biochemical signatures and results from classical assays of element function are complex and modestly predictive in probabilistic ways rather than strongly predictive in a deterministic fashion. Conventional definitions of element functions (e.g. enhancers, silencers, insulators, non-coding RNAs etc) are proving woefully inadequate and incomplete in the face of massive numbers of functional elements that defy simple classification based on sequence or other easily measured features. And, while the relevance of ENCODE elements to disease-associated variation from GWAS studies is exciting, a definitive approach for connecting human genetic and epigenetic variation to disease contexts remains to be realized.

It is now clear that the next phase of functional genomics research will leverage and integrate emerging technological, computational and biological strategies to tackle complex biological problems such as cell differentiation and the etiology of disease. High-throughput approaches for mapping genomic features (biochemical and otherwise) will be complemented by new tools for high-throughput genome engineering and systematic functional perturbation, thus enabling expanded and in some cases qualitatively different approaches to large-scale genome science.

ENCODE is positioned to make an enabling contribution to this broader effort, focusing on areas where the coordinated action of a consortium and large-scale data generation can have the most impact. This contribution must continue to provide high value in an environment where major assay formats (e.g. conventional ChIP-seq) that were once the province of a few well-equipped, high-throughput laboratories have now become widely adopted; where emerging technologies and approaches for genome engineering are undergoing rapid development and dissemination; and where increasingly sophisticated computational tools are becoming more accessible to diverse investigators.

The challenges facing the broader field of functional genomics, and the potential points at which ENCODE can make meaningful enabling contributions can be organized hierarchically into a set of layered goals that encompass distinct experimental and informational components:

Layer 1: ***Completing the Catalog of Elements***

While substantial progress has been made, it is clear that discovery/delineation of functional elements in the human genome is still incomplete. It is also clear that the activity of the vast majority of functional elements is cell context-specific, and that expansion of the catalog of elements will require systematic efforts to characterize and penetrate:

- *New cell and tissue types*. The human body comprises over 400 recognized cell types based on classical microscopic and histochemical modes of analysis; the true number is potentially higher, perhaps significantly so.

- *New types of elements*. The genome encodes diverse functional and physical interactions that are poorly understood (e.g., with 1000s of regulatory factors that bind DNA or RNA)
- *Condition-specific elements*. Many elements are activated in response to particular external stimuli (e.g., steroid response elements) or intrinsic programs such as differentiation.

The vast biological element space will also require implementation of a new generation of mapping/discovery tools that are capable of:

- *Substantially higher sample throughput* (>10X over current platforms), while maintaining high cost efficiency
- *Routine application to small numbers of cells* (500-50,000 cell range) to enable penetration of diverse biologically meaningful compartments

Critically, the above must be achieved without erosion in resolution or data quality compared with current gold-standard assays.

Consortium-driven efforts in these directions could offer significant benefits and efficiencies. Additionally, the expertise and enablement of the broader community has brought new potential for synergy with the consortium toward the goal of creating a large encyclopedia of functional elements. Specific opportunities include:

- *Creating a community-focused data coordination center* to augment and expand consortium efforts by assembling, curating and making easily publicly accessible the high-quality data and corresponding metadata generated by diverse expert community investigators.
- *Creating a truly global resource* by systematically integrating data from other large-scale functional genomics projects (e.g., GGR, GTEx, Epigenomics Consortia) with ENCODE and community data into an easily accessible comprehensive reference.

The above efforts have the potential to make ENCODE data – and those of many other projects ranging from focused R01s to large consortia – more universal, accessible, and useful.

Layer 2: ***Connecting Elements with their cognate gene(s)***

Connecting distal elements with their target gene(s) is vital for maximizing the utility of the Catalog. Achieving this goal will require a highly coordinated effort coupling integrative computational analysis, genome-scale assays, and systematic experimental perturbations that will challenge the limits of high-throughput functional genomics platforms. This type of effort is well suited to a consortium approach, and the nature of the resulting data will be of immediate and ongoing utility for the community.

The challenges encompassed under such an effort are substantial. For example:

- *Different categories of elements will impact different features* – from transcription initiation to elongation to splicing to local and regional chromatin states – many of which may not be readily detectable with conventional assays.
- *Cellular and genomic context sensitivity is likely to be the rule* – individual elements have evolved within a specific chromatin context, and at specific distances from genes and other nearby elements.

- *Many elements are 'primed' or 'memory' sites* – elements that are detectable biochemically (e.g., paused RNA transcripts, certain histone modifications or hypersensitivity) yet impotent within a particular context in which additional activating signals are missing.

The problem of connecting distal elements to their cognate gene(s) has been addressed using several molecular and computational approaches, including:

- *Activity correlation*. The appearance of biochemical signatures at many elements is tightly correlated with the appearance of activating features at the promoters of their cognate gene(s). Because most elements show cell selectivity, analysis of these co-activation patterns over dozens or even hundreds of cell types can systematically connect elements with target genes.
- *Physical interaction.* Many distal elements contact their target promoters (or other elements), which is presumed to be vital for function, and the relative frequencies with which such interactions occur can now be routinely measured with several experimental strategies (e.g., 5C, HiC, ChIA-PET etc.). However, our understanding of how such interactions – or which interactions – are most significant from the functional perspective is still nascent.
- *Knockouts.* Reverse genetics in an isogenic setting is a powerful approach for establishing both function per se, and specific connections between a given DNA segment and control of specific genes.

The last approach is particularly attractive because of its potential to yield definitive answers. And if the readout is chosen to be transcription of the gene (as measured by any number of conventional approaches) the stage is set for systematic analysis of the functional connections between elements and genes – *without requiring detailed knowledge of the precise functional role or contribution of each element* (see below, Layer 3).

Connecting will require integrated experimental and computational tool development to reveal and properly assign physical and regulatory interactions of elements with their target genes.

Layer 3: ***Transforming the Catalog of Elements into a full-fledged Encyclopedia – Categorizing sequence elements into functional behavioral classes***

It is now clear that the human genome encodes a very large number of DNA elements that are marked with biochemical signatures characteristic of important biological functions, and it is obvious that deep functional characterization (under-emphasized in prior stages of ENCODE) will be essential for transforming the catalog of elements – i.e., where functional information is encoded in the genome – into an encyclopedia, wherein each entry describes not only the where, but also the what and how of each element.

The elements defined by ENCODE (and related projects) are both extremely numerous – numbering in the millions – and astonishingly diverse with respect to their (i) sequence features, (ii) patterns of cellular detection, (iii) patterns of factor occupancy, (iv) surrounding chromatin modifications, and (v) broader chromatin structural context.

Transforming the catalog into a full-fledged encyclopedia will thus require systematic categorization of functional elements. We must move beyond the simple assay-driven vocabulary inherited from the 1980s 'Golden Age' of regulation – enhancer, promoter, silencer,

insulator – and fully flesh out the major categories of functional elements encoded by the genome. This challenge is daunting for several reasons

- We currently have little basis for estimating how many such categories may exist
- Many elements are likely to have subtle and complex functions that will only be revealed by integration of multiple data types
- Additionally, many elements may express their functions in a highly context-specific fashion – or even may express different functions in different cellular contexts.

These shortcomings may be addressed with emerging technologies, including (but not limited to) high-throughput synthetic biology and reporter screens, and genome engineering.

The sheer scale and diversity of the problem is not well suited to a highly centralized consortium-style approach. However, ENCODE is well positioned to make an enabling contribution by continuing to develop approaches for computationally categorizing elements (e.g., by chromatin state) and systematically probing these computational classifications with focused application of high-throughput assays with well-defined functional readouts.

Layer 4: ***From general to specific: individual variation in sequence elements and its impact on quantitative phenotypes and disease***

An ultimate goal of functional genomics research is to advance understanding of individual variation, disease susceptibility and mechanism, and thus further progress towards genomic and personalized medicine.

However, it remains immensely challenging to interpret individual sequence variation (or moreso non-sequence-based variation). Non-coding variants tend to have subtle effects, which makes them far more challenging to interpret than knockouts. In the absence of a semi-comprehensive catalog, and without a far more comprehensive understanding of the underlying rules, we typically cannot predict the effect of an individual variant or even identify a readout to look for. Moreover, the small effect sizes mean that large sample sizes will be required – in most cases, beyond the scope of what current and horizon technologies can parse.

Here we anticipate that ENCODE can continue to play an enabling role, in which catalog and analysis tools can aid investigators in their selection of likely functional variants, while efforts and experimental and computational technology development can hasten progress towards the realization of necessary high-throughput and robust tools.

In summary, an overarching goal of ENCODE is to enable the biomedical research community. A future iteration of the ENCODE Project has the potential to take functional annotation of the human genome in health and disease to a new level, if it were armed with a new generation of functional genomic technologies that, by virtue of increased throughput and substantially decreased sample requirements, could be applied systematically to pertinent biological models.

Appendix 6: Acknowledgements

National Advisory Council for Human Genome Research (http://www.genome.gov/10000905)
ENCODE External Consultants Panel (http://www.genome.gov/12513392)
Workshop Organization Committee: Aviv Regev, Carol Bult, Eric Boerwinkle, John Lis.
NHGRI Leadership: Eric Green, Jeffery A. Schloss.
Elise Feingold, Daniel Gilchrist, Mike Pazin, Julie Coursen, Hannah Naughton, Adam Felsenfeld, Alvaro Encinas, Kiara Palmer