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.