Decoding ENCODE

The ENCODE project provides fundamental insights into the genome and large-scale science, inspiring future collaboration at the genomics–chemical biology interface.

The Encyclopedia of DNA Elements (ENCODE) Project (http://www.encodeproject.org/) was initiated by the United States National Human Genome Project Research Institute to create a 'user manual' for the human genome by providing high-quality functional annotations of human genome sequences. Since 2003, the international consortium comprising 32 institutes and 442 members has populated a publicly accessible database (http://genome.ucsc.edu/ENCODE/) with genome annotations derived from systematic profiling of 147 cell lines across 24 standardized experimental platforms. The coordinated publication of 30 research papers by ENCODE researchers in September of this year represents a leap forward in our scientific understanding of genomes, but it also offers insights into scientific collaboration, publishing innovation and research frontiers in chemical biology.

The ENCODE project presents a view of the human genome that has evolved substantially since the Human Genome Project first brought it into focus (*Nature* **489**, 46–48, 2012; *Nature* **489**, 57–74, 2012). The revelation that most of the genome of any given cell is actively transcribed suggests that a more operational definition of a 'gene'—as a dynamic ensemble of RNA transcripts rather than a static modular DNA segment—may be more accurate. Related ENCODE results further compel scientists to begin thinking of the genome as a three-dimensional entity. ENCODE investigators have assigned a functional role to approximately 80% of genomic sequences, most of which fall outside of annotated protein-coding regions and instead lie in domains once referred to as 'junk' DNA. Interestingly, most of the single-nucleotide polymorphisms associated with diseases map to these noncoding regions, reminding us of how much we have to learn about *cis* regulatory sites in DNA, the myriad proteins and RNAs that engage them and the regulatory outcomes of these interactions. These data imply that achieving a mechanistic view of these complex networks will require broader integration of other 'omic' data sets with molecular studies.

Beyond its scientific insights, the ENCODE project provides a useful case study for successful scientific collaborations. Though aspects of the ENCODE project have been criticized, few dispute the importance of the consortium's achievements, and many commend the project's leadership (see, for example, http://bit.ly/REMkqL). In a Comment piece accompanying the *Nature* articles, Ewan Birney, the lead analysis coordinator for ENCODE, outlined the organizational features, leadership structures and investigator commitments that guided the ENCODE team and discussed the lessons learned in the course of the project (*Nature* **489**, 49–51, 2012). With enhanced momentum in the direction of big science, collaboration and data integration in chemical biology (*Nat. Chem. Biol.* **6**, 787–789, 2010; *Nat. Chem. Biol.* **7**, 321, 2011), these insights are likely to be broadly useful to the field as chemical biologists become more engaged in collaborative and data-intensive projects.

Beyond collaborative models, data-rich projects such as ENCODE provide impetus to develop tools for scientists to interact more effectively with large data sets and to integrate results and conclusions across related research studies. The ENCODE explorer site (http://www.nature.com/encode/) created by *Nature*, in collaboration with ENCODE researchers and Illumina, offers one example of such a tool. The explorer presents 13 'threads' on genomic themes that allow users to bring together content relevant to these topical areas from across the ENCODE papers published in *Nature*, *Genome Biology* and *Genome Research*. The supplementary information of the *Nature* papers also contains new functionality that enables users to interact with and analyze ENCODE data directly. These efforts provide important steps toward a more interactive and integrated scientific literature. We encourage chemical biologists to explore the threads and other functionality at the ENCODE explorer site and let us know how similar publishing tools could be useful to the chemical biology community.

The ENCODE project brings scientists one step closer to bridging the gap between genomes and a molecular-level understanding of genes and gene regulation. Though the biological models emerging from the ENCODE project are tantalizing, their mechanistic validation will require substantial interdisciplinary effort. Chemical biologists, starting at the chemical level, have been progressively analyzing systems of increasing biological complexity and should continue to contribute their molecular perspective and tools toward a deeper understanding of the genome. For example, chemical approaches were essential for profiling 5-methylcytosine (5mC) modifications in the ENCODE project. Yet since the inception of ENCODE, it has become clear that numerous 5mC metabolites are present in genomes and that dynamic RNA modification is an emerging regulatory pathway. New platforms for the detection of nucleotide modifications in cellular DNA and in RNA will be essential for a complete understanding of genomic nucleotide metabolism, a challenge that chemical biologists are uniquely equipped to tackle.

Chemical probes will also enable genomic and epigenomic research. The development of chemical inhibitors of chromatin-modifying enzymes is already an active area of chemical biology research (*Nat. Chem. Biol.* **8**, 417–427, 2012) that has facilitated a deeper understanding of the role of epigenetic enzymes in gene regulation. For example, a paper published in the current issue (Article, p. 890; News & Views, p. 875), reports a small-molecule inhibitor of the EZH2 histone methyltransferase and demonstrates its utility as a tool for analyzing the enzyme's role in certain leukemias.

The ENCODE project will similarly point us toward new therapeutic approaches by expanding the druggable genome (*Nat. Rev. Drug Disc.* **1**, 727–730, 2002). In our view, gene regulatory elements, including DNA sequences, transcription factors and noncoding RNAs, have been widely characterized as 'undruggable' targets, mostly because we have an incomplete molecular understanding of these complex systems. We are confident that genomic data, focused mechanistic attention on these systems and chemical tools for the selective manipulation of their components will open future opportunities in genome- and epigenome-targeted therapeutics.

Chemical biologists are poised to advance our understanding of genes and their regulation through future collaboration and broader integration of chemical biology methods and data with omic sciences. Indeed, we anticipate the next edition of the genome's 'user manual' to reflect the necessary convergence of the synthetic and analytical perspectives of chemical biologists with the comprehensive and expansive views of systems biologists. ■