MILESTONE 14

Our most elemental encyclopaedia

Long before the first draft of the human genome was completed, it was well established that the DNA sequence alone is insufficient to control the molecular processes of life. Such information needs to be interpreted and regulated, and protein interactions, chromatin structure and chemical modifications were known to play a key role in this process. At the time, such regulatory processes had only been defined for a limited number of genes. The availability of the blueprint of human genetic information (MILESTONE 1), together with exciting new technologies to analyse gene expression and protein–DNA interactions (MILESTONES 4, 7) formed the foundation for the birth of the Encyclopedia of DNA Elements (ENCODE) project.

For the pilot phase of the project, 35 groups around the world produced and analysed 200 datasets, largely based on microarrays, that comprehensively characterized the functional elements of 30 Mb of the human genome — roughly 1% of the total sequence. Intriguing insights of this initial exploration were that the genome is pervasively transcribed and that many of the newly characterized transcripts did not encode any protein and emerged from regions that were thought to be transcriptionally silent. Epigenetic elements associated with active transcription or DNA replication were identified and sorted into large-scale domains. In addition, functional elements were related to evolutionary constraints and genetic variation, which are critical to understanding both the conservation and adaptability of regulatory processes.

> ENCODE 2 was a landmark for the understanding of molecular genetics and a major feat in data standardization, analysis and integration

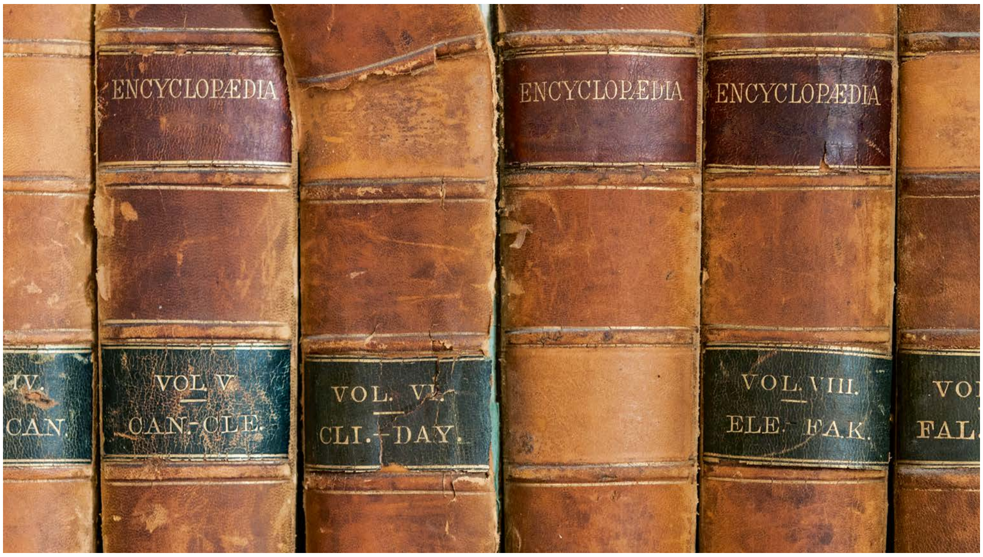
The second version of the encyclopaedia, now applying next-generation sequencing technologies to the entire genome, defined not only a set of 20,687 protein-coding genes but also how their expression is controlled in 147 different cell types. About 80.4% of the genome was associated with at least one biochemical event (that is, RNA- or chromatin-related), and 95% was within reach of a DNA–protein interaction. Technologies to probe long-range physical interactions between distinct chromosomes revealed a plethora of promoter–enhancer interactions that are critical for gene activation.

ENCODE 2 was a landmark for the understanding of molecular genetics and a major feat in data standardization, analysis and integration. Predictive models of gene expression were developed based on epigenetic marks or transcription factor binding patterns. Machine learning methods were trained to cluster functional regions across the genome that were afterwards associated with concrete biological processes, such as immune response or neural activity. There was enough statistical power to quantify the impact of negative selection on genomic regulation, and the integration with large databases of annotated variants enabled the exploration of individual functional alterations, with implications for diseases such as cancer. Meanwhile, modENCODE (fly and worm) and mouse ENCODE began the mapping of functional elements in quintessential model organisms.

Despite the unprecedented resource that the 2012 release of ENCODE represented, the work to characterize the entire functional genome was far from complete. Most of the data had been generated in cell lines, which cannot fully recapitulate the profiles of primary tissues, and most transcription factors had not been assigned to their corresponding genomic elements. The latest version of the encyclopaedia is a collection of 5,992 experiments; it considerably expanded the landscape of regulatory elements both in humans and in the mouse. The ENCODE 3 release also provided maps of cell type-specific 3D chromatin interactions and RNA-binding proteins in human samples, as well as comprehensive profiles of epigenetic changes throughout mouse fetal development.

In the future, we can expect the ENCODE project to expand towards even more comprehensive and functionally tested biochemical profiles, likely incorporating the information from individual genomes and single-cell multi-omics, ensuring that ENCODE remains a crucial reference for our understanding of human biology, evolution and disease.

Ilse Valtierra, *Nature Communications*



Credit: Oleksandra Korobova / Getty images

ORIGINAL ARTICLES Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012) | Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007) | Moore, J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020)