News & views

Genomics

ENCODE expanded

Chung-Chau Hon & Piero Carninci

The third phase of the Encyclopedia of DNA Elements (ENCODE) project has generated the most comprehensive catalogue yet of the functional elements that regulate our genes. **See p.699, p.711, p.720, p.729, p.737, p.744, p.752 & p.760**

Less than 2% of the human genome encodes proteins[1]. A grand challenge for genomic sciences has been mapping the functional elements — the regions that determine the extent to which genes are expressed — in the remaining 98% of our DNA. The Encyclopedia of DNA Elements (ENCODE) project, among other large collaborative efforts[2–4], was established in 2003 to create a catalogue of these functional elements and to outline their roles in regulating gene expression. In nine papers in *Nature*[5–13], the ENCODE consortium delivers the third phase of its valuable project.

In 2007, the pilot phase of the ENCODE project searched for functional elements in 1% of the genome in a few human cell lines[14]. The consortium catalogued two types of these elements. First, they identified DNA regions that are transcribed into RNA (both protein-coding and non-protein-coding). Second, they identified DNA regions that regulate gene transcription, known as *cis*-regulatory elements (CREs). These regions can be identified by their accessibility to DNA-cleaving enzymes such as DNase I, by DNA-binding proteins such as transcription factors, or by specific molecular modifications on histone proteins, to which DNA is bound in a complex called chromatin.

In 2012, the second phase of the ENCODE project extended the search for these functional elements to the whole genome in more human cell lines[15], laying down a solid foundation for the encyclopedia. Similar efforts were extended to the mouse genome in 2014, deepening our understanding of these elements from an evolutionary perspective[16].

In the current third phase of the project, the consortium moved from cell lines to cells taken directly from human and mouse tissues, providing a more biologically relevant encyclopedia. They also introduced assays to investigate the broader aspects of functional elements — for example, to characterize the elements embedded in RNAs or to analyse chromatin looping, which brings separate CREs into close proximity to enable gene regulation (Fig. 1).

In the flagship article, The ENCODE Project Consortium *et al.*[5] (page 699) provide a bird's-eye view of the updated encyclopedia, which contains newly added data sets from 6,000 experiments, performed on around 1,300 samples. By integrating these data sets, the consortium has created an online registry of candidate CREs. Most are classified as promoters or enhancers — CREs respectively located at or some distance from the genomic site at which transcription of a gene begins. The consortium tracked the activity of each candidate CRE, along with the proteins that bind to it in many different samples from various tissues. They used chromatin-looping data to link enhancers to genes that they might regulate. This online registry marks a true milestone, turning an overwhelming amount of genomic information into a searchable, filterable and retrievable encyclopedia of DNA elements, which is freely accessible at https://screen.encodeproject.org.

The eight companion articles, along with a Perspective[17] in this issue and papers in several of *Nature*'s sister journals (see go.nature.com/encode), delve deep into the biology behind this project. These studies leveraged the scale and variety of the ENCODE data sets to reveal the principles that govern how functional elements work. Together, the papers demonstrate the value of large-scale data production in biology.

To identify CREs, Meuleman and colleagues[6] defined 3.6 million accessible DNA regions in 438 cell and tissue types and states, by measuring the genome-wide susceptibility to DNase I. Vierstra and colleagues[7] (page 729) further examined the patterns of DNase I cleavage to understand how transcription factors bind to the CREs. They concluded that most CREs are occupied by multiple transcription factors in an independent and well-spaced manner.

To further understand how transcription factors cooperate at CREs, Partridge and colleagues[8] (page 720) mapped the binding of 208 proteins (including 171 transcription factors) across the genome of a human

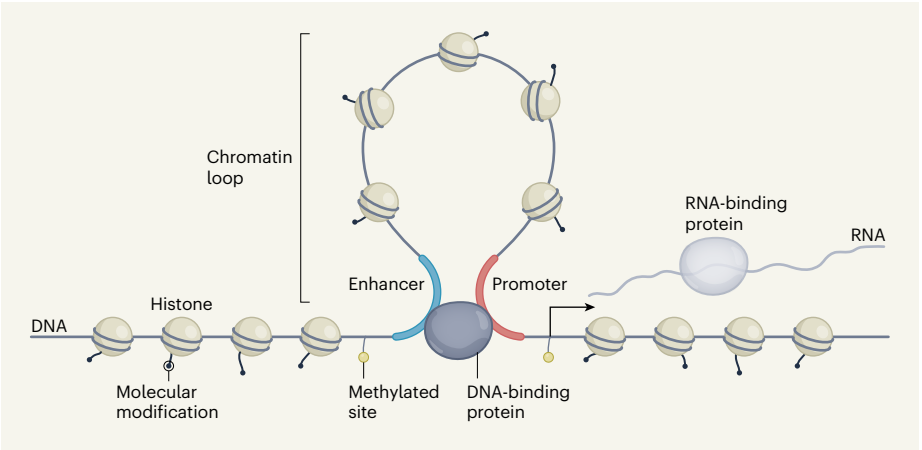

Figure 1 | Examining functional elements across the genome. Nine papers[5–13] from the Encyclopedia of DNA Elements (ENCODE) project catalogue a broad range of aspects of genome organization and regulation. The consortium produced a catalogue of *cis*-regulatory elements (CREs) — DNA sequences called promoters and enhancers that regulate the transcription (black arrow) of genes at close or distant genomic sites, respectively. The researchers analysed DNA and RNA sequences, and the proteins bound to them, including histone proteins, around which DNA is packaged as chromatin. They also studied molecular modifications to DNA and histones, including sites of DNA methylation. These modifications can lead to changes in gene regulation, and in chromatin looping, which can bring enhancers into close contact with the genes they regulate.

liver cell line. This is nearly one-quarter of the proteins that bind to chromatin in this cell line – an unprecedented level of coverage. The map highlights about 5,000 highly occupied target (HOT) regions, mostly promoters and enhancers, that are bound by many transcription factors.

Highly occupied target regions have been described before[18], but – by analysing the patterns in which proteins co-assemble at these regions, and the DNA sequences to which they bind – Partridge *et al.* have provided the first comprehensive evidence to support a speculative model of HOT-region formation[18]. Under this model, a set of anchor DNA sequences first recruit specific transcription factors. These proteins increase chromatin accessibility, then serve as a core around which other binding proteins aggregate in a manner that is independent of DNA sequence. This could happen through protein–protein interactions and chromatin loops, which might link together multiple distant CREs.

To further understand how distant CREs work together, Grubert *et al.*[9] (page 737) mapped chromatin loops in 24 human cell types. They showed that differences in chromatin looping between cell types can affect gene expression, by changing which distant enhancer elements regulate a gene, and which sections of a gene are retained after transcription (a process called alternative splicing). Their most intriguing finding was that housekeeping genes (those involved in day-to-day cell maintenance) often interact with just a few enhancer elements, whereas many enhancers make contact with genes that cause disease if one of two copies is mutated. This implies that a simpler circuitry favours steady and constant expression, whereas more-complex circuitry is needed to safeguard the expression of 'dosage-sensitive' genes.

Three other papers explore the regulatory roles of CREs from a developmental perspective by analysing multiple tissues from mouse fetuses across several developmental stages. Yupeng He *et al.*[10] (page 752) investigated the patterns of DNA modification by methyl groups; Gorkin *et al.*[11] (page 744) studied histone modification and the accessibility of chromatin at whole-tissue levels; and Peng He *et al.*[12] (page 760) analysed gene-expression profiles at the single-cell level.

The first two of these studies[10,11] together reveal a general principle of developmental gene regulation: over time, there is a continuous removal of methyl marks that promote stable gene silencing to enable rapid, flexible modes of gene regulation, controlled by histone modifications and chromatin accessibility. Intriguingly, both groups demonstrated that the human equivalents of some mouse enhancer elements – those that are active only in specific tissues – are enriched for genetic variants associated with diseases relevant to the corresponding tissue. This observation highlights the value of studying CREs in animals.

In the third study[12], Peng He and colleagues harnessed the single-cell resolution of their gene-expression data to computationally 'decompose' the whole-tissue results from the first two studies. In this way, they could predict which enhancer elements are active in specific cell types in a tissue. This study is a perfect demonstration of how integrative analyses enhance the value of data from multiple studies.

Finally, Van Nostrand *et al.*[13] (page 711) present a comprehensive analysis of functional RNA elements by cleverly integrating multiple assays. The authors analysed the RNA sequences that RNA-binding proteins adhere to *in vitro*. They then used these data to interpret the results of another assay that identifies the RNA sequences bound by RNA-binding proteins *in vivo*. To investigate the effects of RNA-binding proteins on genes (in terms of expression level and alternative splicing patterns), they integrated their binding information with assays of the gene-expression profiles of cells in which particular RNA-binding proteins were inhibited. This revealed potential roles for some RNA elements in RNA stability and alternative splicing.

"This encyclopedia has already become a quintessential tool for understanding gene regulation and genetic predisposition to disease."

One surprising finding was that about half of the RNA-binding proteins profiled interact with DNA as well as RNA. However, the authors found little evidence that the proteins bind to both types of molecule at sites that have the same sequence. This suggests that the interactions of most RNA-binding proteins with DNA are not mediated by RNAs bound to chromatin at their site of synthesis. Further studies are needed to assess whether these interactions occur through direct binding, are mediated by other DNA-binding proteins, or are mediated by RNAs bound to chromatin outside their site of synthesis.

This catalogue of RNA elements substantially expands our knowledge of the regulatory components encoded in the human genome. It should enable researchers to predict genetic variants that alter RNA processing, and will constitute an invaluable resource for research into how protein–RNA interactions are regulated.

The third phase of the ENCODE project is a tour de force. But because many regulatory elements act only in specific cell types or at particular times, it is not possible to precisely assess the completeness of the encyclopedia. It would be interesting to see how the project might incorporate single-cell technologies to tease out spatio-temporal-specific elements and so to further unveil the fundamentals of gene regulation. We would also be eager to see the convergence of the ENCODE project with collaborative efforts that have overlapping scopes, for example the Human Cell Atlas[19] (which aims to map gene expression in all the cells of the human body) and the 4D Nucleome project[20] (which aims to understand the 3D organization of the chromosome at different times and in various cell types). The integration of relevant data from these projects into the encyclopedia would further widen its coverage.

This yet-to-be-completed encyclopedia has already become a quintessential tool for understanding gene regulation and genetic predisposition to disease. In the upcoming fourth phase of the ENCODE project, we would be excited to see a systematic assessment of whether the catalogued CREs actually performed the functions inferred on the basis of histone modifications and bound proteins; this could be achieved using high-throughput functional-genomic technologies. The continued expansion of ENCODE to wider biological contexts (such as disease samples and rare cell types) at single-cell resolution would help researchers to use genomic information to diagnose and prevent diseases.

Chung-Chau Hon and **Piero Carninci** are at the RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan.
e-mail: carninci@riken.jp

1. International Human Genome Sequencing Consortium. *Nature* **431**, 931–945 (2004).
2. Roadmap Epigenomics Consortium *et al. Nature* **518**, 317–330 (2015).
3. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. *Cell* **167**, 1145–1149 (2016).
4. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). *Nature* **507**, 462–470 (2014).
5. The ENCODE Project Consortium *et al. Nature* **583**, 699–710 (2020).
6. Meuleman, W. *et al. Nature* https://doi.org/10.1038/s41586-020-2559-3 (2020).
7. Vierstra, J. *et al. Nature* **583**, 729–736 (2020).
8. Partridge, E. C. *et al. Nature* **583**, 720–728 (2020).
9. Grubert, F. *et al. Nature* **583**, 737–743 (2020).
10. He, Y. *et al. Nature* **583**, 752–759 (2020).
11. Gorkin, D. U. *et al. Nature* **583**, 744–751 (2020).
12. He, P. *et al. Nature* **583**, 760–767 (2020).
13. Van Nostrand, E. L. *et al. Nature* **583**, 711–719 (2020).
14. The ENCODE Project Consortium. *Nature* **447**, 799–816 (2007).
15. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
16. Yue, F. *et al. Nature* **515**, 355–364 (2014).
17. The ENCODE Project Consortium *et al. Nature* **583**, 693–698 (2020).
18. Yip, K. Y. *et al. Genome Biol.* **13**, R48 (2012).
19. Regev, A. *et al. eLife* **6**, 27041 (2017).
20. Dekker, J. *et al. Nature* **549**, 219–226 (2017).