Nature Reviews Genetics | AOP, published online 7 September 2012; doi:10.1038/nrg3329

GENOMICS

Users' guide to the human genome

Bananastock

How is the information that is contained in the 3 billion letters of the human genetic code unlocked within individual cells? Answering this question was the task assigned to ENCODE — the Encyclopedia of DNA Elements Consortium — which has now published its analyses in six *Nature* papers and more than two dozen companion papers. Thanks to their collective efforts, there is now at least one biochemical activity assigned to 80% of the human genome.

The ENCODE project was set up to reveal how genetic instructions are read on a global, genome-wide scale and to develop new approaches to allow us to delve into the problem deeper. The Consortium's main findings and their integration are presented in an overview paper. In it, more than 400 authors describe the production and initial analysis of the data and integrate results from diverse experiments with existing resources. What emerges is a detailed picture of human genome organization that has important implications for its function. For example, newly defined chromatin states are identified as hallmarks of genomic regions with distinct functional properties. Numerous non-coding regions are now annotated as a result of the ENCODE effort — 95% of the genome is now known to lie within 8 kb of a DNA–protein interaction. This information is informative to both basic and disease-related human biology.

Five companion papers in *Nature* explore the key themes of ENCODE in more detail.

Gingeras and colleagues chart the transcriptional landscape of the human genome and find that three-quarters of it can be transcribed. For countless known and previously unannotated RNAs, they provide information on the range and levels of expression, localization, processing fates, modifications and associated regulatory regions.

The principles of the human transcriptional regulatory network were investigated by Gerstein *et al.*, who studied the genome-wide binding of 119 transcription factors. They inferred a hierarchical network of transcription factor interactions and found that different parts of the network show different properties: for example, they found that the hierarchy contains a middle tier of transcription factors that introduce information-flow bottlenecks.

DNaseI footprinting was used by Neph *et al.* to study regulatory factor binding sites in 41 cell and tissue types. Collectively, they defined 8.4 million distinct short sequence elements, doubling the number of known human *cis*-regulatory regions. Among them are novel conserved recognition motifs that are cell-type-specific and that behave like previously described major regulators of development, differentiation and pluripotency.

Novel relationships between chromatin accessibility, transcription, DNA methylation and regulatory factor occupancy were revealed by Thurman *et al.* They made a map of human DNaseI hypersensitive sites for 125 cell and tissue types, which they integrated with other ENCODE data sets. Among their findings was an unexpected link between chromatin accessibility, proliferative potential and patterns of human variation.

Dekker and colleagues focused on three-dimensional looping interactions between transcription start sites and distal elements using a high-throughput method known as chromosome conformation capture carbon copy (5C). Generally, they found that genomic proximity is not a good predictor of long-range interactions, as only 7% of looping interactions are with the nearest gene. The three-dimensional interaction landscape turns out to be very complex: distal and proximal regulatory regions engage in multiple long-range interactions and form complex networks.

> "What emerges is a detailed picture of human genome organization that has important implications for its function"

Collectively, the papers describe 1,640 data sets that have been generated across 147 different cell types using an impressive selection of state-of-the-art technologies. Only a subset of the key themes that emerge from these analyses could have been chosen as a main focus for individual papers. Not wanting other important themes to go unnoticed, the Consortium has come up with the Threads — an innovative way of navigating through the results and discussions online.

What comes next after ENCODE? Do we now understand the genome? Clearly not. Investigations of more cell types and transcription factors and a greater emphasis on primary tissues are needed, but above all there is much more analysis and integration to be done. Much of this task will be left to the wider scientific community, and here the ENCODE project has amply provided in terms of the data and analysis tools.

Magdalena Skipper,
Senior Editor, Nature

ORIGINAL RESEARCH PAPERS The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012) | Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012) | Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012) | Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012) | Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012) | Sanyal, A. *et al.* The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012)
FURTHER INFORMATION *Nature* ENCODE page: http://www.nature.com/ENCODE