FORUM: Genomics

ENCODE explained

The Encyclopedia of DNA Elements (ENCODE) project dishes up a hearty banquet of data that illuminate the roles of the functional elements of the human genome. Here, five scientists describe the project and discuss how the data are influencing research directions across many fields. See Articles p.57, p.75, p.83, p.91, p.101 & Letter p.109

Serving up a genome feast

JOSEPH R. ECKER

Starting with a list of simple ingredients and blending them in the precise amounts needed to prepare a gourmet meal is a challenging task. In many respects, this task is analogous to the goal of the ENCODE project[1], the recent progress of which is described in this issue[2–7]. The project aims to fully describe the list of common ingredients (functional elements) that make up the human genome (Fig. 1). When mixed in the right proportions, these ingredients constitute the information needed to build all the types of cells, body organs and, ultimately, an entire person from a single genome.

The ENCODE pilot project[8] focused on just 1% of the genome — a mere appetizer — and its results hinted that the list of human genes was incomplete. Although there was scepticism about the feasibility of scaling up the project to the entire genome and to many hundreds of cell types, recent advances in low-cost, rapid DNA-sequencing technology radically changed that view[9]. Now the ENCODE consortium presents a menu of 1,640 genome-wide data sets prepared from 147 cell types, providing a six-course serving of papers in *Nature*, along with many companion publications in other journals.

One of the more remarkable findings described in the consortium's 'entrée' paper (page 57)[2] is that 80% of the genome contains elements linked to biochemical functions, dispatching the widely held view that the human genome is mostly 'junk DNA'. The authors report that the space between genes is filled with enhancers (regulatory DNA elements), promoters (the sites at which DNA's transcription into RNA is initiated) and numerous previously overlooked regions that encode RNA transcripts that are not translated into proteins but might have regulatory roles. Of note, these results show that many DNA variants previously correlated with certain diseases lie within or very near non-coding functional DNA elements, providing new leads for linking genetic variation and disease.

The five companion articles[3–7] dish up diverse sets of genome-wide data regarding the mapping of transcribed regions, DNA binding of regulatory proteins (transcription factors) and the structure and modifications of chromatin (the association of DNA and proteins that makes up chromosomes), among other delicacies.

Djebali and colleagues[3] (page 101) describe ultra-deep sequencing of RNAs prepared from many different cell lines and from specific compartments within the cells. They conclude that about 75% of the genome is transcribed at some point in some cells, and that genes are highly interlaced with overlapping transcripts that are synthesized from both DNA strands. These findings force a rethink of the definition of a gene and of the minimum unit of heredity.

Moving on to the second and third courses, Thurman *et al.*[4] and Neph *et al.*[5] (pages 75 and 83) have prepared two tasty chromatin-related treats. Both studies are based on the DNase I hypersensitivity assay, which detects genomic regions at which enzyme access to, and subsequent cleavage of, DNA is unobstructed by chromatin proteins. The authors identified cell-specific patterns of DNase I hypersensitive sites that show remarkable concordance with experimentally determined and computationally predicted binding sites of transcription factors. Moreover, they have doubled the number of known recognition sequences for DNA-binding proteins in the human genome, and have revealed a 50-base-pair 'footprint' that is present in thousands of promoters[5].

The next course, provided by Gerstein and colleagues[6] (page 91) examines the principles behind the wiring of transcription-factor networks. In addition to assigning relatively simple functions to genome elements (such as 'protein X binds to DNA element Y'), this study attempts to clarify the hierarchies of transcription factors and how the intertwined networks arise.

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

Beyond the linear organization of genes and transcripts on chromosomes lies a more complex (and still poorly understood) network of chromosome loops and twists through which promoters and more distal elements, such as enhancers, can communicate their regulatory information to each other. In the final course of the ENCODE genome feast, Sanyal and colleagues[7] (page 109) map more than 1,000 of these long-range signals in each cell type. Their findings begin to overturn the long-held (and probably oversimplified) prediction that the regulation of a gene is dominated by its proximity to the closest regulatory elements.

> ***"These findings force a rethink of the definition of a gene and of the minimum unit of heredity."***

One of the major future challenges for ENCODE (and similarly ambitious projects) will be to capture the dynamic aspects of gene regulation. Most assays provide a single snapshot of cellular regulatory events, whereas a time series capturing how such processes change is preferable. Additionally, the examination of large batches of cells — as required for the current assays — may present too simplified a view of the underlying regulatory complexity, because individual cells in a batch (despite being genetically identical) can sometimes behave in different ways. The development of new technologies aimed at the simultaneous capture of multiple data types, along with their regulatory dynamics in single cells, would help to tackle these issues.

A further challenge is identifying how the genomic ingredients are combined to assemble the gene networks and biochemical pathways that carry out complex functions, such as cell-to-cell communication, which enable organs and tissues to develop. An even greater challenge will be to use the rapidly growing body

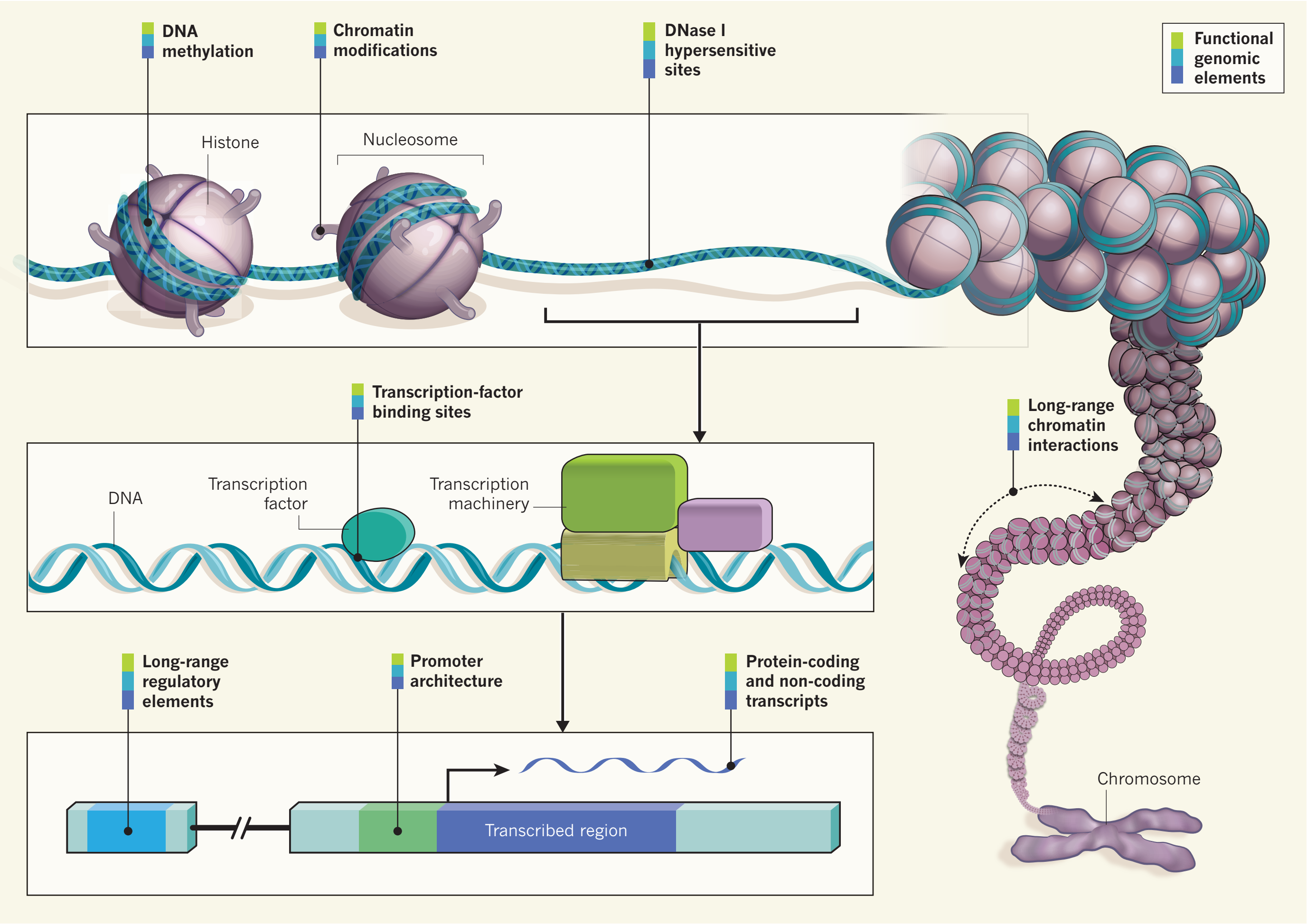
DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
DNA
Transcription factor
Transcription machinery
Long-range chromatin interactions
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

Figure 1 | Beyond the sequence. The ENCODE project[2–7] provides information on the human genome far beyond that contained within the DNA sequence — it describes the functional genomic elements that orchestrate the development and function of a human. The project contains data about the degree of DNA methylation and chemical modifications to histones that can influence the rate of transcription of DNA into RNA molecules (histones are the proteins around which DNA is wound to form chromatin). ENCODE also examines long-range chromatin interactions, such as looping, that alter the relative proximities of different chromosomal regions in three dimensions and also affect transcription. Furthermore, the project describes the binding activity of transcription-factor proteins and the architecture (location and sequence) of gene-regulatory DNA elements, which include the promoter region upstream of the point at which transcription of an RNA molecule begins, and more distant (long-range) regulatory elements. Another section of the project was devoted to testing the accessibility of the genome to the DNA-cleavage protein DNase I. These accessible regions, called DNase I hypersensitive sites, are thought to indicate specific sequences at which the binding of transcription factors and transcription-machinery proteins has caused nucleosome displacement. In addition, ENCODE catalogues the sequences and quantities of RNA transcripts, from both non-coding and protein-coding regions.

of data from genome-sequencing projects to understand the range of human phenotypes (traits), from normal developmental processes, such as ageing, to disorders such as Alzheimer's disease[10].

Achieving these ambitious goals may require a parallel investment of functional studies using simpler organisms — for example, of the type that might be found scampering around the floor, snatching up crumbs in the chefs' kitchen. All in all, however, the ENCODE project has served up an all-you-can-eat feast of genomic data that we will be digesting for some time. Bon appétit!

Joseph R. Ecker *is at the Howard Hughes Medical Institute and the Salk Institute for Biological Studies, La Jolla, California 92037, USA.*
e-mail: ecker@salk.edu

Expression control

WENDY A. BICKMORE

Once the human genome had been sequenced, it became apparent that an encyclopaedic knowledge of chromatin organization would be needed if we were to understand how gene expression is regulated. The ENCODE project goes a long way to achieving this goal and highlights the pivotal role of transcription factors in sculpting the chromatin landscape.

Although some of the analyses largely confirm conclusions from previous smaller-scale studies, this treasure trove of genome-wide data provides fresh insight into regulatory pathways and identifies prodigious numbers of regulatory elements. This is particularly so for Thurman and colleagues' data[4] regarding DNase I hypersensitive sites (DHSs) and for Gerstein and colleagues' results[6] concerning DNA binding of transcription factors. DHSs are genomic regions that are accessible to enzymatic cleavage as a result of the displacement of nucleosomes (the basic units of chromatin) by DNA-binding proteins (Fig. 1). They are the hallmark of cell-type-specific enhancers, which are often located far away from promoters.

The ENCODE papers expose the profusion of DHSs — more than 200,000 per cell type, far outstripping the number of promoters — and their variability between cell types. Through the simultaneous presence in the same cell type of a DHS and a nearby active promoter, the researchers paired half a million enhancers with their probable target genes. But this leaves

11 Years Ago

The draft human genome

OUR GENOME UNVEILED

Unless the human genome contains a lot of genes that are opaque to our computers, it is clear that we do not gain our undoubted complexity over worms and plants by using many more genes. Understanding what does give us our complexity — our enormous behavioural repertoire, ability to produce conscious action, remarkable physical coordination (shared with other vertebrates), precisely tuned alterations in response to external variations of the environment, learning, memory … need I go on? — remains a challenge for the future.

David Baltimore
From *Nature* 15 February 2001

GENOME SPEAK

With the draft in hand, researchers have a new tool for studying the regulatory regions and networks of genes. Comparisons with other genomes should reveal common regulatory elements, and the environments of genes shared with other species may offer insight into function and regulation beyond the level of individual genes. The draft is also a starting point for studies of the three-dimensional packing of the genome into a cell's nucleus. Such packing is likely to influence gene regulation … The human genome lies before us, ready for interpretation.

Peer Bork and Richard Copley
From *Nature* 15 February 2001

more than 2 million putative enhancers without known targets, revealing the enormous expanse of the regulatory genome landscape that is yet to be explored. Chromosome-conformation-capture methods that detect long-range physical associations between distant DNA regions are attempting to bridge this gap. Indeed, Sanyal and colleagues[7] applied these techniques to survey such associations across 1% of the genome.

The ENCODE data start to paint a picture of the logic and architecture of transcriptional networks, in which DNA binding of a few high-affinity transcription factors displaces nucleosomes and creates a DHS, which in turn facilitates the binding of further, lower-affinity factors. The results also support the idea that transcription-factor binding can block DNA methylation (a chemical modification of DNA that affects gene expression), rather than the other way around — which is highly relevant to the interpretation of disease-associated sites of altered DNA methylation[11].

The exquisite cell-type specificity of regulatory elements revealed by the ENCODE studies emphasizes the importance of having appropriate biological material on which to test hypotheses. The researchers have focused their efforts on a set of well-established cell lines, with selected assays extended to some freshly isolated cells. Challenges for the future include following the dynamic changes in the regulatory landscape during specific developmental pathways, and understanding chromatin structure in tissues containing heterogeneous cell populations.

Wendy A. Bickmore *is in the Medical Research Council Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.*
e-mail: wendy.bickmore@igmm.ed.ac.uk

Non-coding but functional

INÊS BARROSO

The vast majority of the human genome does not code for proteins and, until now, did not seem to contain defined gene-regulatory elements. Why evolution would maintain large amounts of 'useless' DNA had remained a mystery, and seemed wasteful. It turns out, however, that there are good reasons to keep this DNA. Results from the ENCODE project[2–8] show that most of these stretches of DNA harbour regions that bind proteins and RNA molecules, bringing these into positions from which they cooperate with each other to regulate the function and level of expression of protein-coding genes. In addition, it seems that widespread transcription from non-coding DNA potentially acts as a reservoir for the creation of new functional molecules, such as regulatory RNAs.

What are the implications of these results for genetic studies of complex human traits and disease? Genome-wide association studies (GWAS), which link variations in DNA sequence with specific traits and diseases, have in recent years become the workhorse of the field, and have identified thousands of DNA variants associated with hundreds of complex traits (such as height) and diseases (such as diabetes). But association is not causality, and identifying those variants that are causally linked to a given disease or trait, and understanding how they exert such influence, has been difficult. Furthermore, most of these associated variants lie in non-coding regions, so their functional effects have remained undefined.

"The results imply that sequencing studies focusing on protein-coding sequences risk missing crucial parts of the genome."

The ENCODE project provides a detailed map of additional functional non-coding units in the human genome, including some that have cell-type-specific activity. In fact, the catalogue contains many more functional non-coding regions than genes. These data show that results of GWAS are typically enriched for variants that lie within such non-coding functional units, sometimes in a cell-type-specific manner that is consistent with certain traits, suggesting that many of these regions could be causally linked to disease. Thus, the project demonstrates that non-coding regions must be considered when interpreting GWAS results, and it provides a strong motivation for reinterpreting previous GWAS findings. Furthermore, these results imply that sequencing studies focusing on protein-coding sequences (the 'exome') risk missing crucial parts of the genome and the ability to identify true causal variants.

However, although the ENCODE catalogues represent a remarkable tour de force, they contain only an initial exploration of the depths of our genome, because many more cell types must yet be investigated. Some of the remaining challenges for scientists searching for causal disease variants lie in: accessing data derived from cell types and tissues relevant to the disease under study; understanding how these functional units affect genes that may be distantly located[7]; and the ability to generalize such results to the entire organism.

Inês Barroso *is at the Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK, and at the University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Cambridge, UK.*
e-mail: ib1@sanger.ac.uk

Evolution and the code

JONATHAN K. PRITCHARD & YOAV GILAD

One of the great challenges in evolutionary biology is to understand how differences in DNA sequence between species determine differences in their phenotypes. Evolutionary change may occur both through changes in protein-coding sequences and through sequence changes that alter gene regulation.

There is growing recognition of the importance of this regulatory evolution, on the basis of numerous specific examples as well as on theoretical grounds. It has been argued that potentially adaptive changes to protein-coding sequences may often be prevented by natural selection because, even if they are beneficial in one cell type or tissue, they may be detrimental elsewhere in the organism. By contrast, because gene-regulatory sequences are frequently associated with temporally and spatially specific gene-expression patterns, changes in these regions may modify the function of only certain cell types at specific times, making it more likely that they will confer an evolutionary advantage[12].

However, until now there has been little information about which genomic regions have regulatory activity. The ENCODE project has provided a first draft of a 'parts list' of these regulatory elements, in a wide range of cell types, and moves us considerably closer to one of the key goals of genomics: understanding the functional roles (if any) of every position in the human genome.

Nonetheless, it will take a great deal of work to identify the critical sequence changes in the newly identified regulatory elements that drive functional differences between humans and other species. There are some precedents for identifying key regulatory differences (see, for example, ref. 13), but ENCODE's improved identification of regulatory elements should greatly accelerate progress in this area. The data may also allow researchers to begin to identify sequence alterations occurring simultaneously in multiple genomic regions, which, when added together, drive phenotypic change — a process called polygenic adaptation[14].

However, despite the progress brought by the ENCODE consortium and other research groups, it remains difficult to discern with confidence which variants in putative regulatory regions will drive functional changes, and what these changes will be. We also still have an incomplete understanding of how regulatory sequences are linked to target genes. Furthermore, the ENCODE project focused mainly on the control of transcription, but many aspects of post-transcriptional regulation, which may also drive evolutionary changes, are yet to be fully explored.

Nonetheless, these are exciting times for studies of the evolution of gene regulation. With such new resources in hand, we can expect to see many more descriptions of adaptive regulatory evolution, and how this has contributed to human evolution.

Jonathan K. Pritchard *and* **Yoav Gilad** *are in the Department of Human Genetics, University of Chicago, Chicago 60637 Illinois, USA.* **J.K.P.** *is also at the Howard Hughes Medical Institute, University of Chicago.*
e-mails: pritch@uchicago.edu; gilad@uchicago.edu

From catalogue to function

ERAN SEGAL

Projects that produce unprecedented amounts of data, such as the human genome project[15] or the ENCODE project, present new computational and data-analysis challenges and have been a major force driving the development of computational methods in genomics. The human genome project produced one bit of information per DNA base pair, and led to advances in algorithms for sequence matching and alignment. By contrast, in its 1,640 genome-wide data sets, ENCODE provides a profile of the accessibility, methylation, transcriptional status, chromatin structure and bound molecules for every base pair. Processing the project's raw data to obtain this functional information has been an immense effort.

For each of the molecular-profiling methods used, the ENCODE researchers devised novel processing algorithms designed to remove outliers and protocol-specific biases, and to ensure the reliability of the derived functional information. These processing pipelines and quality-control measures have been adapted by the research community as the standard for the analysis of such data. The high quality of the functional information they produce is evident from the exquisite detail and accuracy achieved, such as the ability to observe the crystallographic topography of protein–DNA interfaces in DNase I footprints[5], and the observation of more than one-million-fold variation in dynamic range in the concentrations of different RNA transcripts[3].

> ***"The high quality of the functional information produced is evident from the exquisite detail and accuracy achieved."***

But beyond these individual methods for data processing, the profound biological insights of ENCODE undoubtedly come from computational approaches that integrated multiple data types. For example, by combining data on DNA methylation, DNA accessibility and transcription-factor expression. Thurman *et al.*[4] provide fascinating insight into the causal role of DNA methylation in gene silencing. They find that transcription-factor binding sites are, on average, less frequently methylated in cell types that express those transcription factors, suggesting that binding-site methylation often results from a passive mechanism that methylates sites not bound by transcription factors.

Despite the extensive functional information provided by ENCODE, we are still far from the ultimate goal of understanding the function of the genome in every cell of every person, and across time within the same person. Even if the throughput rate of the ENCODE profiling methods increases dramatically, it is clear that brute-force measurement of this vast space is not feasible. Rather, we must move on from descriptive and correlative computational analyses, and work towards deriving quantitative models that integrate the relevant protein, RNA and chromatin components. We must then describe how these components interact with each other, how they bind the genome and how these binding events regulate transcription.

If successful, such models will be able to predict the genome's function at times and in settings that have not been directly measured. By allowing us to determine which assumptions regarding the physical interactions of the system lead to models that better explain measured patterns, the ENCODE data provide an invaluable opportunity to address this next immense computational challenge. ■

Eran Segal *is in the Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel.*
e-mail: eran.segal@weizmann.ac.il

1. The ENCODE Project Consortium *Science* **306,** 636–640 (2004).
2. The ENCODE Project Consortium *Nature* **489,** 57–74 (2012).
3. Djebali, S. *et al. Nature* **489,** 101–108 (2012).
4. Thurman, R. E. *et al. Nature* **489,** 75–82 (2012).
5. Neph, S. *et al. Nature* **489,** 83–90 (2012).
6. Gerstein, M. B. *et al. Nature* **489,** 91–100 (2012).
7. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. *Nature* **489,** 109–113 (2012).
8. Birney, E. *et al. Nature* **447,** 799–816 (2007).
9. Mardis, E. R. *Nature* **470,** 198–203 (2011).
10. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. *Annu. Rev. Med.* **63,** 35–61 (2012).
11. Sproul, D. *et al. Proc. Natl Acad. Sci. USA* **108,** 4364–4369 (2011).
12. Carroll, S. B. *Cell* **134,** 25–36 (2008).
13. Prabhakar, S. *et al. Science* **321,** 1346–1350 (2008).
14. Pritchard, J. K., Pickrell, J. K. & Coop, G. *Curr. Biol.* **20,** R208–R215 (2010).
15. Lander, E. S. *et al. Nature* **409,** 860–921 (2001).