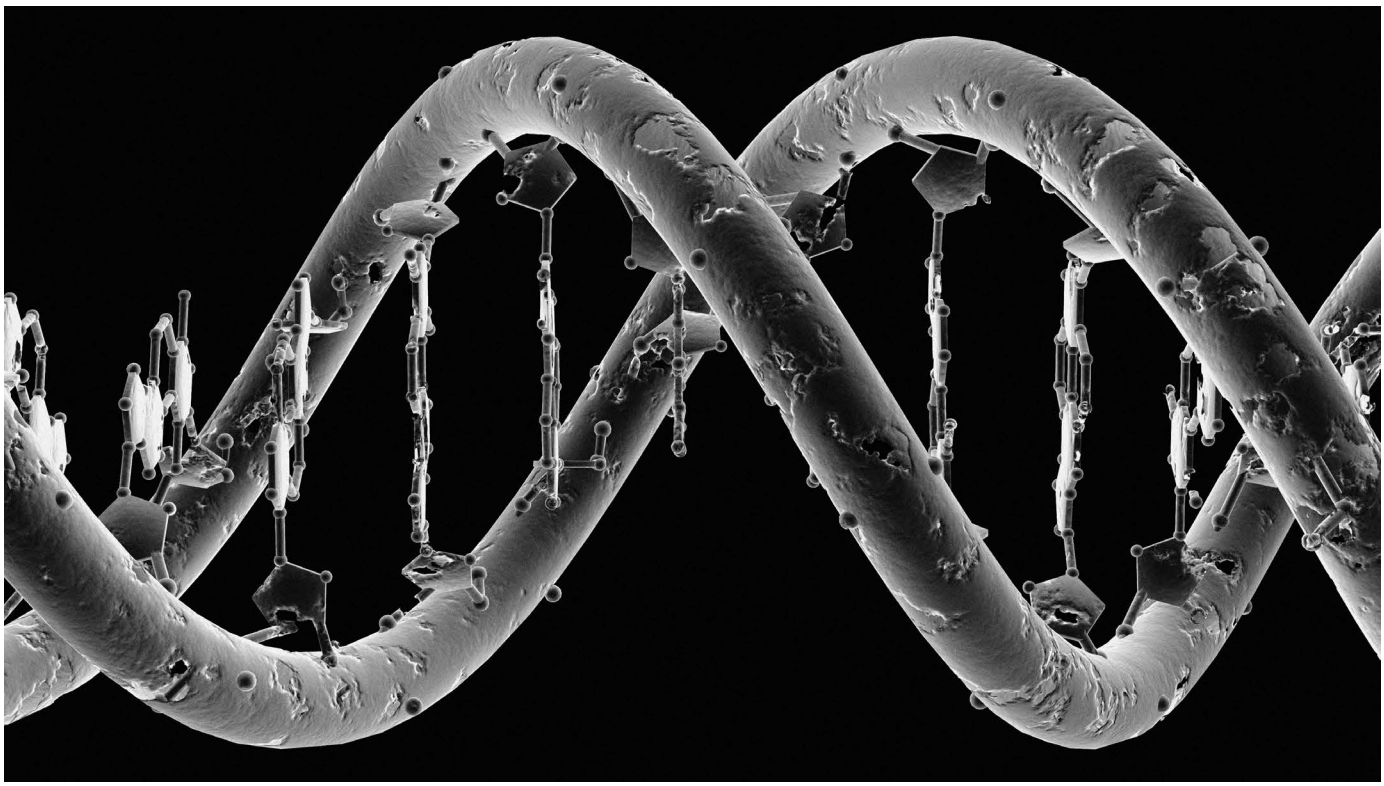
TECHNOLOGY FEATURE

THE DARK SIDE OF THE HUMAN GENOME

Scientists are uncovering the hidden switches in our genome that dial gene expression up and down, but much work lies ahead to peel back the many layers of regulation.

MEHAU KULYK/SPL



The human genome is not packed with 'junk' as previously thought, but with regulatory regions that modulate gene activity.

BY KELLY RAE CHI

Fifteen years ago, scientists celebrated the first draft of the sequenced human genome. At the time, they predicted that humans had between 25,000 and 40,000 genes that code for proteins. That estimate has continued to fall. Humans actually seem to have as few as 19,000 such genes[1] — a mere 1–2% of the genome. The key to our complexity lies in how these genes are regulated by the remaining 99% of our DNA, known as the genome's 'dark matter'.

From efforts such as the massive Encyclopedia of DNA Elements (ENCODE) project[2], launched in 2003 by the US National Human Genome Research Institute, it's clear that copious regulatory elements are at play, tuning gene expression in ways that scientists are only starting to unravel. By uncovering regulatory instructions in the genome beyond protein-coding genes, scientists are hoping to yield new ways to understand and treat disease. "It's not overstating to say that ENCODE is as significant for our understanding of the human genome as the original DNA sequencing of the human genome," says cell biologist Bing Ren of the University of California, San Diego, Institute for Genomic Medicine in La Jolla, who is a member of the ENCODE team.

Ren is also part of a subsequent consortium called the Roadmap Epigenomics Project[3]. These two initiatives — both funded by the US National Institutes of Health (NIH) — aim to map and predict the existence of elements in the genome, including in the vast stretches of non-coding portions, that drive when and where genes are expressed. Scientists have generated a list of such elements by using biochemical assays to probe DNA sequences, RNA transcripts, regulatory proteins bound to DNA and RNA and epigenetic signatures — the chemical tags on DNA and the proteins packaging it — that also affect gene expression.

So far, the data suggest that there are hundreds of thousands of functional regions in the human genome whose task is to control gene expression: it turns out that much more space in the human genome is devoted to regulating genes than to the genes themselves. Scientists are now trying to validate each predicted element experimentally to ascertain its function — a mammoth task, but one for which they now have a powerful new tool. ▶

Since the gene-editing technique CRISPR–Cas9 entered the scientific arena, the speed at which researchers can test functional elements in the non-coding regions has ramped up. But it is still a daunting endeavour: more than 3 million regulatory DNA regions, thought to contain some 15 million binding sites for regulatory proteins called transcription factors, control gene expression in the human cell types studied thus far. About 150,000 may be active in any given cell type.

These could be crucial to understanding disease, because most single-nucleotide changes associated with common diseases fall in regions outside protein-coding genes, and they often overlap with DNA sites highlighted by ENCODE as having regulatory function. Certain regulatory elements that normally drive gene expression are thought to underpin the mechanism of cancer, for example. Disrupting a gene's regulatory elements, the data suggest, could thus have as drastic an impact on cell function as disrupting the gene itself. Using CRISPR–Cas9, scientists now have an opportunity to test that premise by introducing targeted mutations into non-coding sequences and observing the consequences.

DECODING A COMPLEX WORLD

How much of DNA's dark matter has a function in gene control is still up for debate. In 2012, ENCODE scientists proposed on the basis of biochemical-assay predictions that 80% of the non-coding genome has a function[2]. But this figure soon proved to be an overestimate as researchers narrowed the definition of 'function' and devised experimental methods, such as reporter assays, to test these functions. "The number still isn't fully known", in part because the mapping isn't complete, says Michael Snyder, a geneticist at Stanford University in California and a member of ENCODE. "Most people would say between 10% and 20% of the [non-coding] genome is likely to have a function where, if you disrupt it, you will affect something."

But regulatory elements have a bewildering array of functions and forms, which makes tackling them a formidable challenge. Even the best-known types, such as spots in the genome known as promoters, which lie next to a gene where transcription begins, and enhancers — regions that when bound by specific transcription factors alter the likelihood of a gene being read — are hard to study. In addition to the sheer number of these sites, estimated at 15 million, enhancers may be positioned thousands of base pairs away from the gene that they control. This makes it tough to predict where their target genes are located and what they do.

Thus far, ENCODE and Roadmap have offered up important clues, but the real proof that these predicted regulatory elements actually do something comes from a functional test. For genes, this mostly entails deleting them one at a time and observing the consequences in a cell assay or animal model. This is less easy to do for the non-coding genome because many of the elements are redundant, and so deleting just one might not alter gene expression or produce an obvious change. "It's a huge challenge that we have at the moment to really distinguish between functional and non-functional elements detected by ENCODE," says geneticist Ran Elkon of Tel Aviv University in Israel.

CRISPR–Cas9 is particularly accelerating scientists' exploration of enhancers. The technology enables scientists to alter large numbers of regulatory elements in a high-throughput way, using libraries of RNA guide fragments that target and disrupt different regions in the genome, to observe the outcome. Not only is the method relatively fast, but researchers can also run the assays directly in human cells.

Experiments of this type have already turned up some unexpected findings. While a postdoc working with cancer biologist Reuven Agami at the Netherlands Cancer Institute in Amsterdam, Elkon performed the first screen of regulatory elements using the advanced editing system[4]. The CRISPR–Cas9 approach enabled them to test individually those enhancers predicted by ENCODE to bind a transcription factor called p53. Interest in p53 is high because the protein is a known tumour suppressor that is mutated in more than 50% of human tumours. The researchers were able to pinpoint two enhancers from more than a thousand genomic sites that affect p53's tumour-suppressing function, located near the p53-encoding gene. A predicted third enhancer has yet to be located because it is far from any gene, let alone one related to p53.

In a separate screen, the group targeted binding sites for oestrogen receptor-α — which is implicated in breast cancer — and identified three enhancer sequences that influence tumour growth; these elements could thus have a role in the development of resistance to breast-cancer therapy.

"It's a huge challenge to distinguish between functional and non-functional elements."

At the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, bioengineer Feng Zhang and his group also used CRISPR–Cas9 to identify genes essential to the survival of cancer cells. Using a melanoma model, they first screened around 18,000 genes in human cells to pinpoint ones that might underlie resistance to the melanoma drug vemurafenib. Then, in a follow-up study published last month[5], they described a new screen that identified regulatory regions on either side of several resistance genes. Their findings fit well with ENCODE data that predict regulatory regions at these locations — and they also reveal new functional elements, says molecular biologist Neville Sanjana, who conducted the research as a former postdoc in Zhang's group and now works at the New York Genome Center at New York University.

Other CRISPR–Cas9 screening data have challenged ENCODE predictions. Richard Sherwood of Harvard Medical School in Boston, Massachusetts, and his collaborators created an approach called a multiplexed editing regulatory assay[6] to screen for non-coding regions that might influence gene expression in well-known mouse embryonic stem-cell lines. Using this technique, they obtained quantitative information about the extent to which these regulatory regions might contribute to gene expression. Some of their results are discordant with regions flagged by ENCODE as potential enhancers because, when mutated, these areas did not affect gene expression.

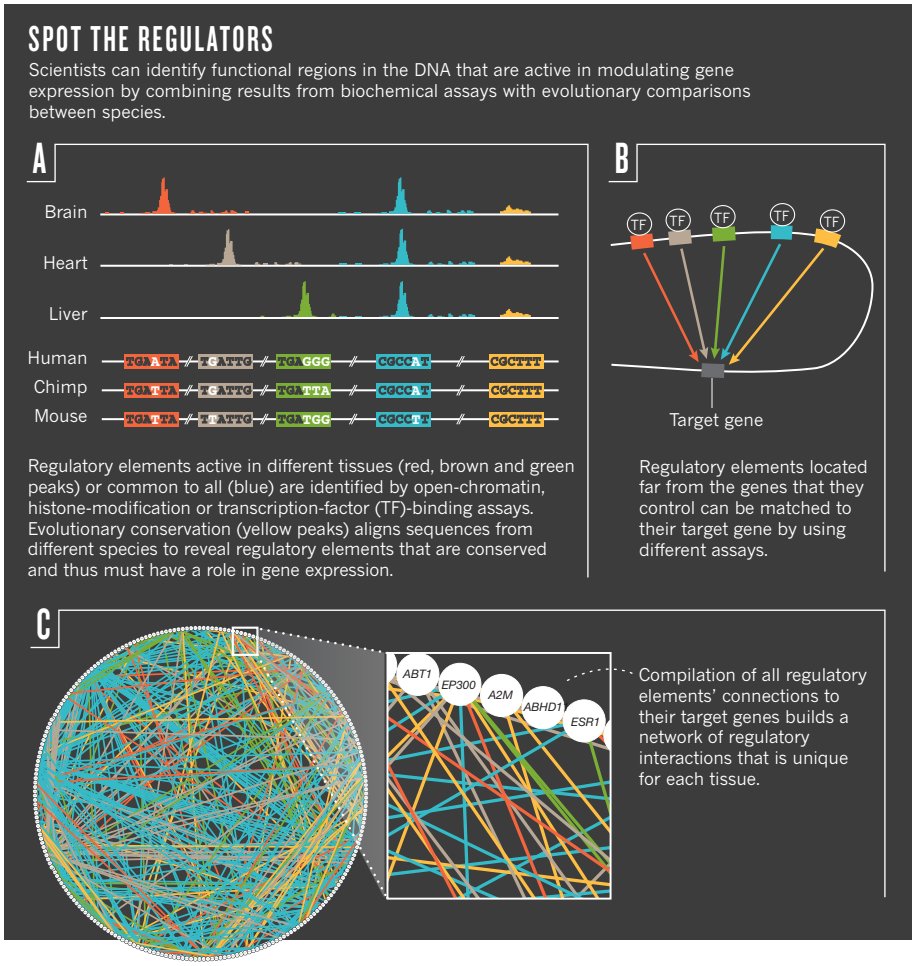
Moreover, the researchers also discovered mysterious sections that they dubbed 'unmarked regulatory elements', or UREs, that do not fit into any category of functional elements. The team is currently exploring how widespread these UREs might be in the genome. This new type of assay, along with other gene-editing-based screens, will play an increasingly important part in the validation of ENCODE candidates, says Sherwood.

TECHNIQUE TWEAKS

Investigators working on the ENCODE and Roadmap projects have relied mostly on a biochemical technique called DNase-seq, which sequences and maps all exposed regions of the genome. In these sections, the DNA is relaxed instead of tightly coiled around histones, and thus is more likely to facilitate transcription-factor binding that drives gene activation. By mapping these areas, investigators can pinpoint candidate enhancers, promoters, silencers, insulators and other regulatory elements in the non-coding genome (see 'Spot the regulators').

Another method, ATAC-seq, detects and sequences sites in the chromatin that are accessible to the transposase enzymes used for the assay. Both DNase-seq and ATAC-seq produce a genome-wide view of regions of open chromatin. According to the researchers, because such epigenomic profiles can map the extent to which genes are activated in certain cell types, they could be useful for clinical decision-making, and ATAC-seq is fast enough for this purpose[7]. Many, however, consider a technique known as chromatin immunoprecipitation (ChIP)-seq to be the most reliable for this purpose because it is the only one that can identify all potential binding sites for a given transcription factor.

Even so, biochemical assays can only hint at function. CRISPR–Cas9 cell screens, by contrast, are more concrete because scientists can introduce a mutation or deletion at a particular site in the genome and observe how it influences gene expression. The disadvantage is that these



ADAPTED FROM E. KHURANA *ET AL. NATURE REV. GENET.* **17,** 93–108 (2016).

tests cover smaller portions of the genome. If the full genome of 3 billion base pairs were represented, for example, by three copies of Leo Tolstoy's classic novel *War and Peace* (1869), such screens would barely cover a single page, Sanjana says — although he is optimistic that future gene-editing approaches will scale this up.

"In the short term, I think CRISPR will serve mainly as a tool to validate functions predicted by those biochemical signatures," says Ren. Once enough of these kinds of screens have been done, their data could be fed into a machine-learning tool to improve its predictive power, Sherwood says.

New computational tools are already providing scientists with smart ways to interpret biochemical mapping data. Algorithms can predict transcription-factor binding sites, which researchers can then probe for function. But even with algorithms, predicting which enhancers are active in a given context is harder in human genomes than in yeast or worm genomes, says computational biologist Michael Beer of Johns Hopkins University in Baltimore, Maryland.

Beer and his collaborators have developed a computational model[8] to predict which tissue-specific networks of gene-regulatory elements are operating in a given cell type and to what extent they are perturbed in complex diseases. They trained their open-source algorithm, called deltaSVM, on human lymphoblastoid cell lines using gene data from ENCODE in 2012, followed by mouse ENCODE data in 2014.

Scientists have initially focused on cancer to probe the links between functional elements and disease because cancer is a simpler condition to study at the cell level than, say, a neuropscyhiatric disorder — cancer cell lines reveal simple-to-measure outcomes, such as cell multiplication, death or senescence. But the data that have streamed in from the Epigenome Roadmap consortium are shifting scientists' thinking about how cancers arise. A study published last year by geneticist John Stamatoyannopoulos of the University of Washington in Seattle and his collaborators showed[9] that mutations in a given cancer cell type cluster in inaccessible chromatin regions rather than in the exposed ones — possibly because the open regions can be accessed by DNA-repair enzymes.

The scientists also found that mutation density in a tumour is defined by the epigenomic profile specific to each type of cell. Consequently, the DNA sequence can be informative about tumour origin, which ushers in the possibility of using epigenomic data to trace cancer provenance in patients for whom it remains unknown. It could also open up new approaches to cancer treatment. "Cancer is essentially a regulatory or epigenetic program that is superimposed on a cell, and the result of that program is the development of genetic and genomic instability," says Stamatoyannopoulos. "As we've analysed lots of cancer genomes, all of these patterns now are starting to come out that were previously not imagined to exist."

It's possible that there are still elements in the genome that existing assays have missed. After all, regulatory signals still crop up unexpectedly, such as the UREs in Sherwood's screen. And a team of scientists led by Harvard Medical School immunologist Daniel Tenen discovered[10] a potential new class of regulators that seem to control whether a gene is turned on or off by blocking the enzyme DNA methyltransferase 1, which adds methyl groups to silence genes. These elements are dubbed 'extracoding RNAs', and because they can influence silencing in a gene-specific way, have therapeutic potential. Earlier this year, neuroscientist Jeremy Day of the University of Alabama at Birmingham and his colleagues showed in rat neurons that an extracoding RNA influences the transcription of a gene important for memory formation[11].

The ENCODE team will continue to map the non-coding space in the genome and expects to cover most of the regulatory DNA by 2020, Stamatoyannopoulos says. A spatial understanding of how DNA is packaged into a cell, and of the 3D folding that positions genes in close contact with their regulatory elements, will be key to predicting an element's target genes. The NIH Common Fund has begun the '4D Nucleome' project, for instance, which aims to predict the target genes for every regulatory element. That knowledge will help to fill in the picture of how a given regulatory element influences health and disease.

Next-generation sequencing has been — and still is — the technological engine of ENCODE. But looking ahead, researchers might be able to roll out high-resolution live-cell imaging on a large scale to watch the state of the genome change in real time using specific markers. This technology could be disruptive. "If we had a better microscope, we wouldn't be sequencing anymore," says Stamatoyannopoulos. ■

Kelly Rae Chi *is a freelance science writer based in Cary, North Carolina.*

1. Ezkurdia, I. *et al. Hum. Mol. Genet.* **23,** 5866–5878 (2014).
2. ENCODE Project Consortium *Nature* **489,** 57–74 (2012).
3. Bernstein, B. E. *et al. Nature Biotechnol.* **28,** 1045–1048 (2010).
4. Korkmaz, G. *et al. Nature Biotechnol.* **34,** 192–198 (2016).
5. Sanjana, N. E. *et al. Science* **353,** 1545–1549 (2016).
6. Rajagopal, N. *et al. Nature Biotechnol.* **34,** 167–174 (2016).
7. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. *Nature Meth.* **10,** 1213–1218 (2013).
8. Lee, D. *et al. Nature Genet.* **47,** 955–961 (2015).
9. Polak, P. *et al. Nature* **518,** 360–364 (2015).
10. Di Ruscio, A. *et al. Nature* **503,** 371–376 (2013).
11. Savell, K. E. *et al. Nature Commun.* **7,** 12091 (2016).