GENOMICS

Mice in the ENCODE spotlight

Following on from affiliated projects in humans and model invertebrates, the Mouse ENCODE Project presents comprehensive data sets on genome regulation in this key mammalian model. SEE ARTICLES P.355, P.365, P.371 & LETTER P.402

PIERO CARNINCI

The mouse genome was sequenced in 2002 as a primary model in which to study gene function and human diseases and to develop drugs[1]. This was followed by maps of transcribed messenger RNA molecules and of long, non-protein-coding RNAs, which facilitated such experiments and analysis[2]. Yet although 17 mouse strains have been sequenced[3], genome function and regulation cannot be understood by sequence analysis alone. Now, in four papers published in this issue[4–7], the Mouse ENCODE Consortium presents data sets that dramatically enhance our understanding of the regulation of the mouse genome, and of the similarities and differences compared with the human genome.

The ENCODE project[8,9] was started by the National Human Genome Research Institute in 2003, with the aim of mapping functional elements of the human genome. The project, later expanded as Mouse ENCODE and modENCODE (to include invertebrate model organisms), has driven technology development and standardization for the identification of expressed RNAs and regulatory regions. These technologies have given rise to comprehensive data sets for analysing genome regulation and comparing this across species. Among the resources are libraries of mRNA sequences and maps of genomic regions that are bound by transcription factors or by RNA polymerases (the enzymes that initiate RNA transcription). There are also data sets on chemical modifications to the histone proteins around which DNA is wrapped (forming a complex called chromatin). Such modifications alter the accessibility of the DNA to other proteins and thereby demarcate transcriptionally 'active' or 'repressed' chromatin regions. And there are data on large-scale chromatin and chromosome structures.

The Mouse ENCODE Project has taken advantage of the ENCODE experience to provide a much-needed comprehensive resource for mouse genomics and its first in-depth analysis. Stergachis and colleagues' data[5] (page 365) reveal that, in the roughly 75 million years of evolution since humans and mice diverged, the primary (nucleic-acid) sequence of regulatory elements has changed dramatically. About half of the transcription-factor binding sites in regulatory elements of the mouse genome are not present in the equivalent (orthologous) elements in humans, and around one-quarter of them have migrated to different positions (Fig. 1). Regulatory elements that are distant from the gene that they regulate (enhancers) have diverged more than those that are close (promoters). Despite this divergence, Cheng *et al.*[6] (page 371) show that there is similar chromatin activity in orthologous promoter regions in the two genomes, suggesting that different transcription factors could be used to achieve similar transcriptional activity. Furthermore, despite the different primary sequences of many regulatory elements, the basic reciprocal regulatory networks among transcription factors are evolutionarily conserved between mice and humans[5].

Surprisingly, the Mouse ENCODE Consortium (Yue *et al.*[4]; page 355) finds that sequences commonly considered useless or harmful, such as retrotransposon elements (stretches of DNA that have been incorporated into chromosomal sequences following reverse transcription from RNA), have species-specific regulatory activity. Because retrotransposon elements can contain embedded transcription-factor binding sites, this may provide unexpected regulatory plasticity (Fig. 1). Evolutionary conservation of primary sequence is typically considered synonymous with conserved function, but this finding suggests that this concept should be reinterpreted, because insertions of retrotransposon elements in new genomic regions are not conserved between species.

Although gene expression might be expected to be similar in the same organs and tissues in different species, comparative analyses by the consortium[4] reveal that the expression level of many genes (but not all gene categories) is species specific, rather than organ specific. These differences may derive from the fact that organs are composed of different cell types in mouse and human tissues, but it is more likely to have arisen from different basic transcriptional activity driven by different regulatory elements.

Despite these variations between the mouse and human genomes, Cheng *et al.*[6] show that many single-nucleotide sequence differences that have been associated with diseases in genome-wide association studies in humans are localized to orthologous regions of the mouse genome that have modifications that mark active chromatin. This finding validates the importance of the mouse as a model organism for ongoing disease studies.

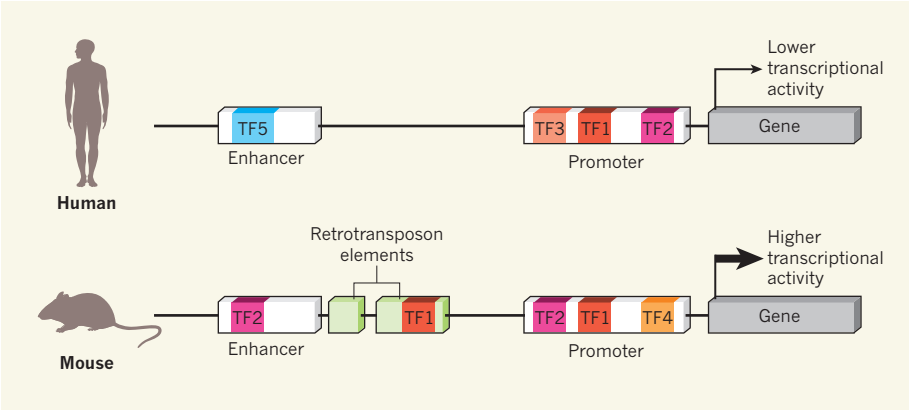

Figure 1 | Transcription-factor binding in mice and humans. Gene transcription rates are regulated by transcription factors, which bind to promoter regions close to the specific gene or to enhancer regions at distant sites. Comparisons of maps of such binding sites generated by the mouse and human ENCODE projects[4–7] suggest that many differences in transcription levels between equivalent (orthologous) genes in the two organisms result from transcription-factor binding sites (labelled as TFs) occupying different locations. A further regulatory influence is the insertion of retrotransposon elements (stretches of DNA derived from reverse transcription of RNA) that may contain transcription-factor binding sites.

Finally, Pope *et al.*[7] (page 402) have generated high-quality maps of the physical position of chromosomes in the nuclei of mouse and human cells. These maps show that the boundaries of replication domains (genomic regions that replicate at the same time during cell division) correlate well with topologically associating domains — chromosome structures that are associated with the regulation of gene expression.

Analysis of these data will continue, both broadly and in the context of specific biological questions, although new tools for visualizing, analysing and interpreting such data are needed to open them up for broader use by experimental biologists. But the existing findings are already thought-provoking. For example, they suggest that we should rethink the relationship between genomic function and evolutionary conservation. Regulatory regions and long non-coding RNAs (lncRNAs) are not subject to the evolutionary constraints of protein-coding genes, which may help to explain the sequence drifts reported in these papers. However, it is striking that transcription-factor networks are conserved despite low conservation of their binding positions in the genome. Further experiments are needed to establish whether transcription-factor interactions with regulated regions always promote transcription or whether they can also be repressive. The differences in regulation between mice and human genomes that have emerged from these studies should all be taken into account when using mouse models to assess biological functions and, in particular, drug responses.

Some genomic features in particular, such as lncRNAs, warrant further investigation. The Mouse ENCODE Project analysed only RNA molecules that are polyadenylated (they have a string of adenine bases at the 3′ end); although this modification marks most mRNAs, many lncRNAs are not polyadenylated[10], and so analysis of non-polyadenylated RNAs in mice will be needed to better define the similarities and differences between the full complement of RNA transcripts in mice and humans. A comprehensive map of orthologous human and mouse lncRNAs will also be useful for experimental tests of the function of human lncRNAs in mice.

Furthermore, there is room to expand the data set on transcription-factor binding sites generated by Cheng and colleagues[6], because their experiments were performed using mouse cells that are easy to cultivate (MEL and CH12) and thus provide plenty of experimental material, but they do not represent the biological variability present in the hundreds of cell types found in mammals[11]. It will also be useful to replicate these studies in different mouse strains and to connect differences in genome sequence[3] between the strains to differences in gene regulation and traits.

The data sets provided by the mouse ENCODE project boost our capacity to analyse the mouse genome in a way that was unthinkable a decade ago, and allows us to gain insights into dimensions that were not foreseeable. Understanding genomic regulation in mice is much more than a linear addition to our knowledge of genome regulation overall — it is an essential step towards better understanding human biology and improving biomedical applications and drug development. ■

Piero Carninci *is at the RIKEN Center for Life Science Technologies, Division of Genomic Technologies, RIKEN Yokohama Campus, Yokohama, Kanagawa 230-0045, Japan.*
e-mail: carninci@riken.jp

1. Chinwalla, A. T. *et al. Nature* **420,** 520–562 (2002).
2. The FANTOM Consortium *et al. Science* **309,** 1559–1563 (2005).
3. Keane, T. M. *et al. Nature* **477,** 289–294 (2011).
4. Yue, F. *et al. Nature* **515,** 355–364 (2014).
5. Stergachis, A. B. *et al. Nature* **515,** 365–370 (2014).
6. Cheng, Y. *et al. Nature* **515,** 371–375 (2014).
7. Pope, B. D. *et al. Nature* **515,** 402–405 (2014).
8. The ENCODE Project Consortium. *Nature* **447,** 799–816 (2007).
9. The ENCODE Project Consortium. *Nature* **489,** 57–74 (2012).
10. Djebali, S. *et al. Nature* **489,**101–108 (2012).
11. The FANTOM Consortium *et al. Nature* **507,** 462–470 (2014).

ORIGINS OF LIFE

RNA made in its own mirror image

An RNA enzyme has been generated that can assemble a mirror-image version of itself. The finding helps to answer a long-standing conundrum about how RNA molecules could have proliferated on prebiotic Earth. SEE LETTER P.440

SANDIP A. SHELKE & JOSEPH A. PICCIRILLI

Many organic and biological molecules come in right-handed and left-handed versions that are mirror-image twins of one another. These variations are referred to as D- and L-enantiomers, respectively. Modern RNA molecules are linear polymers that are synthesized from ribonucleotide monomers, and take the D-form. But on page 440 of this issue, Sczepanski and Joyce[1] suggest that early evolution may have involved an interplay between the D- and L-structures of RNA.

Before DNA and proteins existed, RNA may have evolved as the primordial macromolecule that could both store information like DNA does and catalyse chemical reactions like many proteins do. According to this 'RNA world hypothesis'[2], one of the functions of these RNA enzymes (called ribozymes) was to replicate other RNA molecules by using their sequences as templates to make complementary strands. This function, called

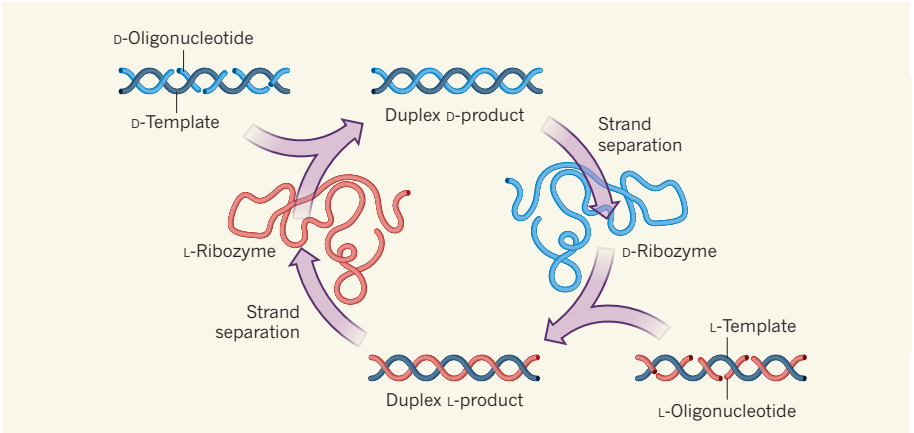

Figure 1 | Possible mechanism for RNA replication on prebiotic Earth. Sczepanski and Joyce[1] have generated an RNA enzyme (a ribozyme) that catalyses the polymerization of oligonucleotides of the opposite handedness to itself: the right-handed D-ribozyme yields the left-handed L-ribozyme, and vice versa. This adds weight to the idea that a cross-handed cycle involving both D- and L-ribozymes may have replicated RNA on prebiotic Earth. In the cycle, the L-ribozyme acts on a complex formed between a D-template RNA strand and D-oligonucleotides, joining the latter together to form a duplex RNA product. Separation of the duplex's strands liberates the D-ribozyme. This then catalyses formation of the L-ribozyme from the left-handed template–oligonucleotide complex.