regulatory regions and the factors that bind to such sequences drive differences in the regulation of Oct4 expression between mouse and cow blastocysts.

It would be interesting to test, in transgenic mice, whether regulatory elements of the human *OCT4* gene behave like the mouse or the cow sequences. Although human blastocysts, like those of domestic animals, express Oct4 in the trophectoderm for an extended period compared with mice, the period of overlap of Cdx2 and Oct4 expression is only slightly longer than in the mouse. Human OCT4 is clearly restricted to the ICM by day 6 before embryo implantation[6].

But why do these regulatory differences exist among the blastocysts of different mammals? Evolutionarily, the placenta is a recent invention, and still seems to be a work in progress. There is huge variation in trophectoderm and placental morphology across different mammalian species, accompanied by recent evolutionary divergence in placenta-specific gene families[7]. For example, a mouse blastocyst attaches and implants in the uterus by embryonic day 5 (E5); a human blastocyst grows a little larger but then implants by E7–9 with highly invasive trophoblast outgrowth; and in cows, pigs and sheep the blastocyst floats in the uterus for 2–3 weeks before attaching.

Berg *et al.* propose that such differences lead to earlier restriction of trophectoderm cell fate in the mouse than in the cow. Indeed, results of their experiments — involving chimaeric blastocysts generated by mixing trophectoderm cells from different stages of development with host embryos — support this proposal.

In a remarkable technical tour de force, they also transferred the chimaeric cow blastocysts to recipient cows and recovered them later in development to show that early trophectoderm cells can contribute to developing ICM derivatives. This is one of the first attempts to test the timing of lineage restriction in a species other than the mouse.

This study emphasizes the need to explore the timing and mechanism of functional lineage restriction in blastocysts of different mammals, including humans. Differences in these parameters may underlie the known difficulty in deriving validated pluripotent embryonic stem cells and trophoblast stem cells from many mammalian species. Although fibroblasts have been reprogrammed into induced pluripotent stem cells in several domestic species, including the cow, these lines often depend on continued expression of exogenous reprogramming factors. Clearly, we need a better understanding of the control of pluripotency in all these species.

As we learn more about the precise details of mouse blastocyst development, we must be constantly evaluating similarities and differences between them and those of humans and other species. This will help us to truly understand mammalian embryo diversity. ■

Janet Rossant *is in the Program in Developmental and Stem Cell Biology, Hospital for Sick Children, and in the Department of Molecular Genetics, University of Toronto, Ontario M5G 1X8, Canada.*
e-mail: janet.rossant@sickkids.ca

1. Berg, D. K. *et al. Dev. Cell* **20,** 244–255 (2011).
2. Nichols, J. *Cell* **95,** 379–391 (1998).
3. Strumpf, D. *et al. Development* **132,** 2093–2102 (2005).
4. Rossant, J. *Reprod. Fertil. Dev.* **19,** 111–118 (2007).
5. Blomberg, L., Hashizume, K. & Viebahn, C. *Reproduction* **135,** 181–195 (2008).
6. Chen, A. E. *et al. Cell Stem Cell* **4,** 103–106 (2009).
7. Wildman, D. E. *Placenta* **32,** 142–145 (2011).

MOLECULAR BIOLOGY

A fly in the face of genomics

The modENCODE project uses integrative analysis to annotate genomic elements in the fruitfly and a nematode worm. The first fly data have now been published. **SEE ARTICLES P.473 & P.480 & LETTER P.527**

EILEEN E. M. FURLONG

The fruitfly *Drosophila melanogaster* is an exceptional model for dissecting the basic principles of biology, development and disease. It is amenable to genetic manipulation using tools developed over more than a century; and its genome shares extensive genetic content with humans. The first draft of the *Drosophila* genome was released a decade ago[1], and with subsequent updates its annotation is in a 'mature' state. Nevertheless, more than half of the predicted genes have been awaiting experimental verification of their structure — the location of promoter sequences, of boundaries of protein-coding and non-coding sequences, and of transcription termini. The modENCODE consortium project aims to address this issue and to identify new genes and genomic elements in the fly genome[2]. Here I focus on the first wave of papers, including three in this issue[3–5], which describes the fly data so far.

To determine which genes are expressed at specific stages of development, Graveley *et al.*[3] (page 473) generated high-resolution expression data, which are complemented by an analysis of 25 *Drosophila* cell lines[6,7]. These efforts identified almost 2,000 new genes that encode proteins or non-coding RNAs. They also extensively refine existing annotation by describing more than 3,000 new promoter sequences[7], roughly 53,000 new or revised exon sequences[3], a threefold increase in RNA-splicing events[3] and a tenfold increase in RNA-editing events[3]. Notably, most of the RNA-editing and -splicing events occur at precise stages of the *Drosophila* life cycle, indicating extensive temporal regulation of these post-transcriptional events by as-yet poorly understood mechanisms. This comprehensive view of the fly transcriptome[3,6,7] reveals that some 75% of the organism's genome is transcribed at one stage or another — in line with the widespread transcription observed in other species.

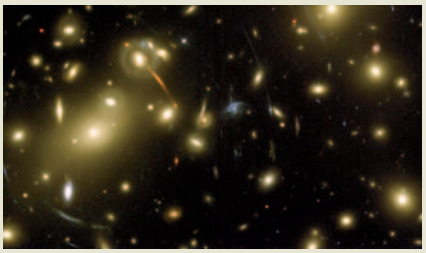
Post-translational histone modifications covering a gene's promoter or coding region provide telltale signatures of the expression status of a gene and thereby present another way to identify functional elements in the genome. Two of the modENCODE studies involved mapping such chromatin marks in *Drosophila* cell lines[4] and at 11 stages of its life cycle[5].

By examining the distribution of 18 histone modifications in two cell lines, Kharchenko *et al.*[4] (page 480) identified nine prominent chromatin signatures, which complement those defined previously[8]. Clues to their function come from information on chromatin accessibility and transcriptional activity, revealing chromatin signatures that distinguish between active and inactive genes, active promoters, and the location of new putative regulatory elements. The authors' global analyses[4] extend previous studies[9–12] indicating that the Polycomb system — a group of chromatin-binding proteins traditionally associated with stable, long-term gene repression during embryonic development — can also function dynamically and associate with promoters that are actively transcribed or seem poised for activation.

Deposition of chromatin marks is linked to the enzymatic activity of RNA polymerase during the initiation and elongation steps of transcription; this activity is regulated by transcription factors bound to *cis*-regulatory elements — proximal and distal sequences that affect gene expression. To understand how transcription is regulated, Nègre *et al.*[5] (page 527) made a systematic effort to identify all *cis*-regulatory elements by examining the occupancy of 38 transcription factors and other chromatin-regulatory proteins at different stages of development. The result is a collection of around 20,000 putative regulatory elements that include insulators, enhancers and

COSMOLOGY

Lenses under the lens



NASA, A. FRUCHTER & THE ERO TEAM [S. BAGGETT (STSCI), R. HOOK (ST-ECF), Z. LEVAY (STSCI)] (STSCI)

If you were to pop into a cosmology conference today, the chances are that you would see this image in at least one presentation. It is a striking snapshot of a cluster of galaxies acting as a gravitational lens: the cluster bends light from galaxies lying behind it and 'smears' the light to produce multiple images and giant arcs.

As pretty as their effects are, gravitational lenses are giving cosmologists a few headaches. For example, the observed incidence of giant arcs and their distance from the clusters' centres, which marks the size of features called Einstein rings, indicate that these clusters may have a stronger 'lensing' ability than expected in the framework of the currently accepted model of the cosmos. In a paper to appear in *Astronomy & Astrophysics*, Meneghetti *et al.* describe an analysis that advances our understanding of these systems (M. Meneghetti *et al.* Preprint at http://arXiv.org/abs/1103.0044; 2011).

The authors compared the lensing ability of a numerically simulated sample of clusters with that of a sample of well-characterized, X-ray-luminous clusters obtained by the MAssive Cluster Survey (MACS). In contrast to earlier studies, their simulations factor in elements known to affect lensing power — for example, the fact that the lenses are complex three-dimensional structures. They found that the simulated clusters produce 50% fewer arcs than do the observed MACS clusters, and that the median size of Einstein rings differs by 25% between the two samples. These are much smaller discrepancies between theory and observation than previously reported. But as the authors themselves concede, more data are needed to confirm their findings. **Ana Lopes**

promoters. Of the more than 2,000 putative promoters, 50% are already confirmed[13]. The locations of about 14,500 putative *cis*-regulatory elements were also identified. Unexpectedly, one class of active promoters does not contain the characteristic chromatin mark H3K4me3, suggesting that the genes they regulate use an alternative mode of transcriptional initiation.

Integrating the binding patterns of all transcription factors leads to hypotheses of transcription-factor partnerships, involving co-binding to regulatory elements[5,7]. But overlays of transcription-factor binding should be interpreted cautiously, particularly for factors with non-tissue-specific or partially overlapping expression: regions that are co-targeted by multiple factors are not necessarily co-bound in the same cells. Nevertheless, the complexity of some co-targeted regions is intriguing. The modENCODE researchers identified regions in the genomes of both *Drosophila*[5,7] and the nematode *Caenorhabditis elegans*[14] — the other model organism on which the project focuses — that are highly occupied by transcription factors. It remains to be determined what function, if any, such regions have in transcription.

This first phase of modENCODE has made a significant impact on refining the annotation of the *Drosophila* genome, which forms the foundation of a large body of research conducted in this organism. But where should the project go from here? First, there is the issue of completion. With the new data, the annotation of genes may be 80% complete, but the job is far from over. Despite the huge depth of coverage, almost 1,500 known genes could not be identified in any experiments[4]. Analysis of specific subpopulations of cells and tighter staging of the developmental process should greatly improve sensitivity.

Completing annotation of the 'regulatory genome' is much more challenging. Although the location of putative enhancer elements can be identified, determining which of these regions are functional, and when, is a huge task. Understanding the regulation of enhancer activity requires knowledge of which transcription factors are binding to them, in which cell types, and when. Scaling this up to the roughly 700 predicted *Drosophila* transcription factors is a monumental undertaking, but feasible given current tagging technologies[15,16].

A major drawback of the data sets is their lack of temporal and spatial resolution. Although cells in culture are extremely useful for identifying core properties of basic cellular processes, such immortalized cells, devoid of their developmental context, cannot substitute for cells within a developing embryo. On the other hand, whole-embryo studies provide merged signals from all cells in the embryo, giving no information on the tissue in which a gene, promoter or chromatin state is active. Many of the transcription factors examined are expressed across a broad range of tissues, which has the advantage of covering a wide range of *cis*-regulatory elements. But merged transcription-factor occupancy signals from multiple tissues make it very difficult to disentangle regulatory connections and thus to build reliable regulatory networks.

The general absence of functional information is perhaps the most serious limitation of the current work and a major challenge for all genomics projects. Such information is essential to understand the relevance of regulatory connections. Examining mutants was understandably beyond the scope of the present studies, but, moving forward, there is a clear need to integrate diverse types of functional data in order to make the transition from correlations to regulatory function. The thousands of *Drosophila* mutants available should provide a useful resource for this.

We can view this work[3–5] as an important chapter in a long book. The data — all freely available[17] — provide an excellent resource for identifying putative genes and regulatory elements that might be active at a particular stage of development. The sheer volume of new transcripts and putative regulatory elements, and the inherent complexity of their interactions, demonstrates how far the project has come, but also highlights the challenges that lie ahead to convert this wealth of information into regulatory networks that describe the transformation of a fertilized egg into a complex multicellular organism. To reach this goal, researchers must integrate new types of experiments that will address the function of, and connections between, genomic regions at high spatio-temporal resolution. With this in mind, we can envisage a next phase of exciting studies that will tackle these issues, and so look forward to seeing what comes next. ■

Eileen E. M. Furlong *is at the Genome Biology Unit, European Molecular Biology Laboratory, D-69117 Heidelberg, Germany.*
e-mail: furlong@embl.de

1. Adams, M. D. *et al. Science* **287,** 2185–2195 (2000).
2. Celniker, S. E. *et al. Nature* **459,** 927–930 (2009).
3. Graveley, B. R. *et al. Nature* **471,** 473–479 (2011).
4. Kharchenko, P. V. *et al. Nature* **471,** 480–485 (2011).
5. Nègre, N. *et al. Nature* **471,** 527–531 (2011).
6. Cherbas, L. *et al. Genome Res.* **21,** 301–314 (2011).
7. The modENCODE Consortium *Science* **330,** 1787–1797 (2010).
8. Filion, G. J. *et al. Cell* **143,** 212–224 (2010).
9. Papp, B. & Müller, J. *Genes Dev.* **20,** 2041–2054 (2006).
10. Kwong, C. *et al. PLoS Genet.* **4,** e1000178 (2008).
11. Oktaba, K. *et al. Dev. Cell* **15,** 877–889 (2008).
12. Enderle, D. *et al. Genome Res.* **21,** 216–226 (2011).
13. Hoskins, R. A. *et al. Genome Res.* **21,** 182–192 (2011).
14. Gerstein, M. B. *et al. Science* **330,** 1775–1787 2010).
15. Venken, K. J. T. *et al. Nature Methods* **6,** 431–434 (2009).
16. Ejsmont, R. K., Sarov, M., Winkler, S., Lipinski, K. A. & Tomancak, P. *Nature Methods* **6,** 435–437 (2009).
17. www.modencode.org