FUNCTIONAL GENOMICS

The modENCODE guide to the genome

Now that obtaining genome sequence is routine, assigning function is the current frontier. Vast data sets that are now available for the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes — and which are described in a raft of new papers — show that large-scale collaborative efforts offer a way forward. This work by the model organism Encyclopedia of DNA Elements (modENCODE) Project offers unprecedented functional annotation and is likely to provide the foundation for countless future experimental and computational studies.

The modENCODE Project was instigated in 2007 (in parallel with an expansion of the human ENCODE Project, the pilot phase of which had analysed 1% of the human genome) with the aim of identifying all sequence-based functional elements in the worm and fruitfly genomes. Two integrative papers that summarize the work and several companion papers have now been published.

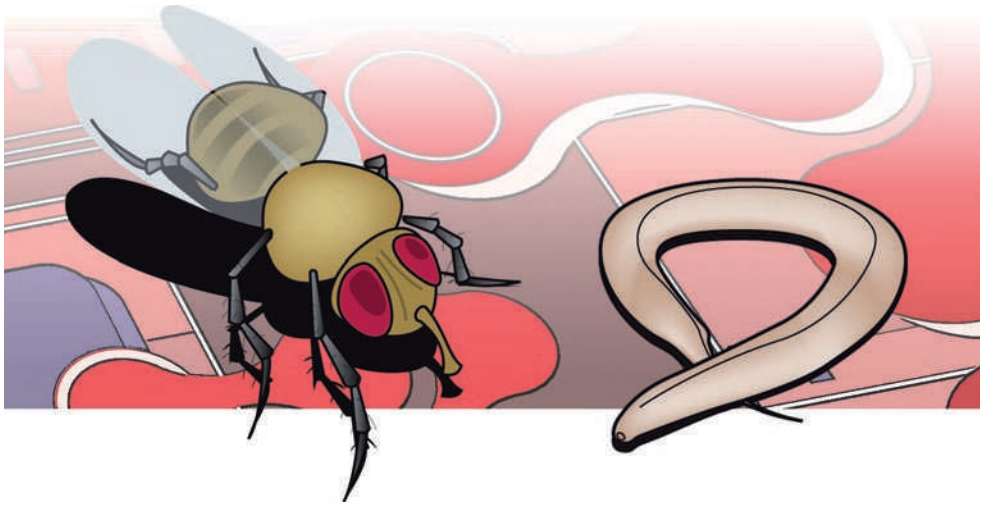
The genome-wide data sets collected — 237 for *C. elegans* and 700 for *D. melanogaster* — include high-throughput RNA sequencing (RNA-seq), chromatin immunoprecipitation followed by sequencing (ChIP–seq) for transcription factor binding and histone modifications, DNA replication patterns and nucleosome occupancy. Importantly, the researchers analysed samples from a range of developmental stages and cell lines, which will help to discern the functional importance and dynamics of DNA features.

Integrative analysis of these data sets was particularly important for studying gene regulation, which by its nature integrates DNA sequence information, transcription factors, RNA and chromatin. For example, studies of worms and flies identified numerous short regions — termed highly occupied target (HOT) regions — that are enriched for the binding of many transcription factors. Intriguingly, their genomic locations suggest functional importance — for example, in *C. elegans* they are enriched near to genes that are highly expressed throughout development, and in *D. melanogaster* they overlap origins of replication. HOT regions have novel sequence motifs that could point towards unidentified modes of cooperative transcription factor recruitment.

Data-set integration also revealed functional regulatory networks that can help to improve predictions of gene function and expression. In *D. melanogaster* these networks were used to make predictions of target gene expression based on the expression of their regulators, and to predict the function of previously unannotated genes. The authors found that one-quarter of genes have predictable expression during embryogenesis and that regulatory models can predict their expression under novel conditions (in cell lines). The 'predictable' genes may be those with more precise regulation, although predictions might be possible for more genes with the addition of further data sets.

This parallel work in *C. elegans* and *D. melanogaster* could highlight shared and distinct features of the two species. Comparison of the two species to each other, and to the human ENCODE project, should enable substantial advances in identifying fundamental regulatory principles across animal genomes and provide a model for understanding the role of DNA sequence in gene regulation and disease.

Mary Muers



ORIGINAL RESEARCH PAPERS Gerstein, M. B. *et al.* Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE Project. *Science* **330**, 1775–1787 (2010) | The modENCODE Consortium *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010)
FURTHER READING Hawkins, R. D., Hon, G. C. & Ren, B. Next-generation genomics: an integrative approach. *Nature Rev. Genet.* **11**, 476–486 (2010)
WEBSITE
The modENCODE Project:
http://www.modencode.org