GENOMICS

Hiding in plain sight

The latest releases from the ENCODE and modENCODE research consortia more than double the number of data sets on functional elements in the worm, fly and human genomes. SEE ARTICLES P.393, P.400 & LETTERS P.445, P.449, P.453

FELIX MUERDTER & ALEXANDER STARK

One of the major scientific achievements of our time has been the sequencing of the human genome and those of model organisms such as fruit flies and worms. These sequences encode species-specific information about protein-coding and non-coding genes and the regulatory information that determines when and where the genes are activated. However, even though this genomic information is present in the sequences, understanding it, or even just comprehensively identifying and annotating the different functional elements, is a major challenge. In an effort to identify all functional elements in the genomes of humans, *Drosophila melanogaster* flies and *Caenorhabditis elegans* worms, the Encyclopedia of DNA Elements (ENCODE) and the Model Organism ENCODE (modENCODE) research projects were launched[1,2]. This issue of *Nature* contains five papers[3–7] that summarize the latest data from these consortia. Together, the publications add more than 1,600 new data sets, bringing the total number of data sets from ENCODE and modENCODE to around 3,300 (Fig. 1).

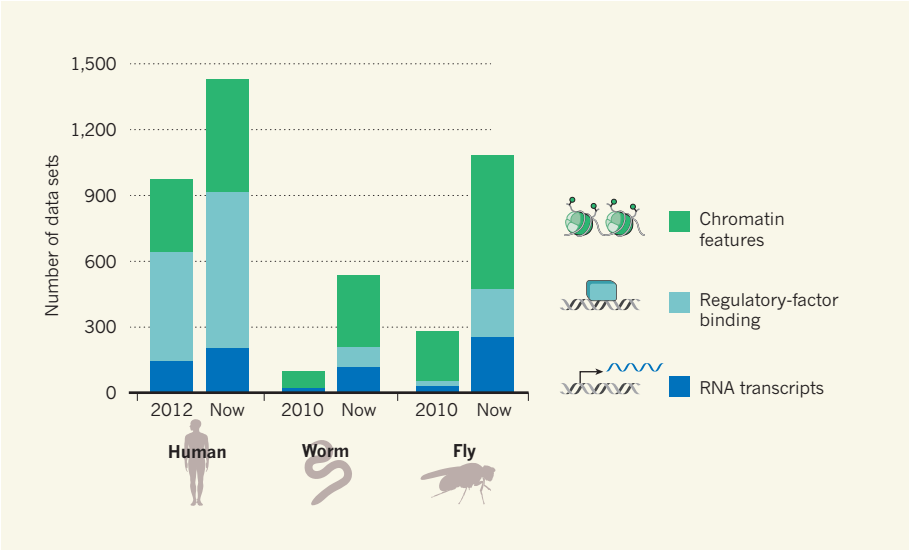

Figure 1 | The growth of ENCODE and modENCODE data sets. The ENCODE and modENCODE research consortia aim to identify all functional elements in the human genome and the genomes of the model organisms *Caenorhabditis elegans* (worm) and *Drosophila melanogaster* (fruit fly). The latest release[3–7] from these projects focuses on three key data types: RNA-seq, which identifies RNA transcripts from cells or whole organisms; ChIP-seq for regulatory factors, which identifies locations in the genome that are bound by these proteins; and sequencing-based assays to profile various features of chromatin (the complex of DNA and histone proteins). The graph shows the total number of data sets now available for these data types, compared with previous releases[19–21] (note that the numbers for the previous worm and fly releases do not include some microarray-based data sets).

The potential impact of such data is undeniable. More-complete genome annotations will form the basis for improved genetic studies in *D. melanogaster* and *C. elegans* — organisms that have already contributed most to our understanding of animal development and the molecular mechanisms involved. It is also increasingly clear that gene-regulatory elements are crucial for development and are frequently linked to disease; comprehensive identification of these elements should, for example, allow the interpretation of disease-associated mutations in non-coding genomic regions.

Two of the papers present data on RNA transcripts — Brown *et al.*[3] (page 393) in *Drosophila* and Gerstein *et al.*[4] (page 445) in all three species. Brown and colleagues' analysis of the *Drosophila* transcriptome, which they assessed in 29 tissues, 24 cell lines and 21 whole-animal samples that had been subjected to environmental perturbations, yielded more than 300,000 transcripts for 17,564 genes, of which 14,692 were protein-coding (different transcripts from the same gene are referred to as transcript isoforms). Of these genes, 57 (5,259 transcripts) were expressed only during perturbations and would thus probably escape identification under standard laboratory conditions. The analysis also identified many new candidate long non-coding RNAs, including ones that overlap with previously defined mutations that have been associated with developmental defects. Another intriguing finding was a small number of mostly neuronal genes that give rise to half of all detected transcript isoforms, reminiscent of the many transcripts known to be generated from the neuronal gene *Dscam*[8]. These data show that sampling selected tissues under non-standard conditions allows new genes and transcript isoforms to be identified even in well-studied organisms.

Regulatory elements are more difficult to identify than transcripts. They are typically predicted on the basis of characteristic features of chromatin (the complex of histone proteins and DNA in the cell nucleus) and by studying regulatory-protein binding to DNA[9] — refining such predictions is a key aim of both the ENCODE and modENCODE projects. Among the latest releases, Araya *et al.*[5] (page 400) report the genome-wide binding profiles for 92 regulatory proteins, including transcription factors, RNA-polymerase subunits and chromatin-associated factors, in whole embryos and larvae from different developmental stages in *C. elegans*. Although this approach may provide information on regulatory changes during development, it is limited by a lack of cellular resolution[10]: transcription factors typically associate with cell-type-specific partner proteins to bind to different sites and regulate distinct genes in different cell types. Therefore, targets that are bound in only a few cells could be missed in whole-organism studies, and those that are found may constitute a superposition of binding sites from different cells. The authors partly deconvoluted these by determining the expression patterns for 180 genes, including

13 of the transcription factors profiled, in the early embryo at single-cell resolution.

Araya and colleagues' data also include binding profiles for predicted transcription factors that are otherwise uncharacterized. This will allow hypotheses to be generated about the proteins' possible functions, particularly, for example, if the binding sites are enriched near certain types of gene[11,12].

A key feature of this rollout of ENCODE and modENCODE data are comparisons across the three species studied. Complementing Araya and colleagues' data in worms, Boyle *et al.*[6] (page 453) present almost 500 new genome-wide binding maps for transcription-regulatory factors in human cell lines, *Drosophila* and *C. elegans*. They found that about half of the binding events in each species occur at high-occupancy target (HOT) regions[13,14], where binding is heavily clustered. Although the function of these regions has not been assessed, our work in *Drosophila*[15] suggests that many are active enhancers, which trigger gene transcription. However, because factors can bind DNA without functional consequences, especially at HOT regions, the contribution of each of the bound factors to enhancer activity remains unclear.

Apart from the existence of HOT regions, Boyle and colleagues' data reveal only a few commonalities between the species. But this is not unexpected — regulatory connections and target genes for individual transcription factors vary substantially between different cell types in a single species, so it is not surprising that there is little overlap in data derived from samples as disparate as human cell lines and whole fly and worm embryos. Thus, although the data sets may be valuable in each of the species, their usefulness for studying the evolution of gene regulation in cross-species comparisons is questionable, because such studies should compare homologous cell types that have shared developmental and functional properties.

Ho and colleagues' comparisons[7] (page 449) focused on chromatin features that characterize regulatory genomic elements, such as DNA accessibility and certain modifications to histone proteins. In 800 new chromatin data sets, they identified several features common to the three species, including shared histone-modification patterns around genes and regulatory regions. Gerstein *et al.* integrated this information with transcription data to present a 'universal model' for predicting gene expression. As the authors point out, these commonalities are not surprising[7] and are in agreement with the modifications' known distributions in each of the three species and in yeast. Instead, Ho and colleagues focused on the observed differences, which predominantly concern chromatin regions that are repressive (gene transcription from such regions is suppressed).

These five papers represent a substantial addition to the public ENCODE and modENCODE resources. We expect the transcriptome data sets to have a direct influence on gene annotations in all three species, which should affect the work of many researchers immediately[16,17]. It is arguably more difficult for scientists to easily access the data on chromatin features and regulatory-factor binding sites, and the regulatory-element predictions. This needs integration with the community portals[16,17] and intuitive interfaces that allow data visualization and flexible analyses, which are being developed by the UCSC Genome Browser project and Ensembl, the two consortia, and others (such as i-cisTarget[11] or GREAT[12]). The success of the ENCODE and modENCODE resources depends on such interfaces being integrated into workflows throughout the research community.

Furthermore, although they are extremely data-rich, the papers expose how data sets that are created to catalogue all functional elements under standardized conditions are not sufficient for understanding the regulation of transcription, chromatin biology and enhancer function, nor the evolution of these mechanisms. Addressing such questions typically requires more-diverse set-ups and experiments, often specifically adjusted for each question. In addition, the identification of regulatory elements remains limited[10] by the lack of cell-type specificity and the fact that chromatin features and regulatory-factor binding are imperfect predictors of regulatory-element function[9]. The papers do not reveal how many of these elements might be functional, and independent estimates span a broad range[9,18]. However, the new data, in conjunction with the work of many other groups, will undoubtedly aid future research into the identification, functional characterization and understanding of genes, regulatory elements and animal genomes more generally. ■

Felix Muerdter *and* **Alexander Stark** *are at the Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), 1030 Vienna, Austria.*
e-mail: stark@starklab.org

1. Celniker, S. E. *et al. Nature* **459,** 927–930 (2009).
2. The ENCODE Project Consortium. *Science* **306,** 636–640 (2004).
3. Brown, J. B. *et al. Nature* **512,** 393–399 (2014).
4. Gerstein, M. B. *et al. Nature* **512,** 445–448 (2014).
5. Araya, C. L. *et al. Nature* **512,** 400–405 (2014).
6. Boyle, A. P. *et al. Nature* **512,** 453–456 (2014).
7. Ho, J. W. K. *et al. Nature* **512,** 449–452 (2014).
8. Schmucker, D. *et al. Cell* **101,** 671–684 (2000).
9. Shlyueva, D., Stampfel, G. & Stark, A. *Nature Rev. Genet.* **15,** 272–286 (2014).
10. Furlong, E. E. M. *Nature* **471,** 458–459 (2011).
11. Herrmann, C., Van de Sande, B., Potier, D. & Aerts, S. *Nucleic Acids Res.* **40,** e114 (2012).
12. McLean, C. Y. *et al. Nature Biotechnol.* **28,** 495–501 (2010).
13. Moorman, C. *et al. Proc. Natl Acad. Sci. USA* **103,** 12027–12032 (2006).
14. The modENCODE Consortium *et al. Science* **330,** 1787–1797 (2010).
15. Kvon, E. Z., Stampfel, G., Yáñez-Cuna, J. O., Dickson, B. J. & Stark, A. *Genes Dev.* **26,** 908–913 (2012).
16. Yook, K. *et al. Nucleic Acids Res.* **40,** D735–D741 (2012).
17. St. Pierre, S. E., Ponting, L., Stefancsik, R., McQuilton, P. & the FlyBase Consortium. *Nucleic Acids Res.* **42,** D780–D788 (2014).
18. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. *Genome Res.* http://dx.doi.org/10.1101/gr.173518.114 (2014).
19. The ENCODE Project Consortium. *Nature* **489,** 57–74 (2012).
20. Gerstein, M. B. *et al. Science* **330,** 1775–1787 (2010).
21. The modENCODE Consortium *et al. Science* **330,** 1787–1797 (2010).

ASTROPHYSICS

Supernova seen through γ-ray eyes

Observations of γ-ray photons from a type Ia supernova indicate that stellar explosions of this kind get their energy from sudden thermonuclear fusion in the progenitor star. SEE LETTER P.406

ROBERT P. KIRSHNER

On page 406 of this issue, Churazov *et al.*[1] report a great discovery — not because it is a surprise, but precisely because it is not. The researchers have detected γ-ray emission lines from the type Ia supernova 2014J in the nearby galaxy M82 using the European Space Agency's INTEGRAL spacecraft. For decades, astronomers have been working out the physical picture for this type of exploding star on the basis of the optical light it emits. The authors' study confirms directly the most fundamental idea in that picture by observing a supernova in the γ-ray range of the electromagnetic spectrum. The γ-rays they observed in the months after the supernova explosion were produced, as expected, by the radioactive decay of isotopes fused in a thermonuclear flame that destroyed a compact star.

Astronomers react quickly to explosive events. On the evening of 21 January 2014, while supervising an undergraduate astronomy