Drosophila genome takes flight

Michael Boutros and Norbert Perrimon

In the March 24 issue of *Science*, a flurry of papers report on the impending completion of the *Drosophila melanogaster* genome sequence. This historic achievement is the result of a unique collaboration between the Berkeley *Drosophila* Genome Project (BDGP), led by Gerry Rubin, and the genomics company Celera, headed by Craig Venter. With its genome almost completely sequenced ahead of schedule, *Drosophila* is another important model organism to enter the postgenomic age, and represents the largest genome sequenced to date.

For almost nine decades, studies of *Drosophila* have been central to our general understanding of genome organization. The wealth of genetic data on mutant phenotypes and the advance of technologies to manipulate the *Drosophila* genome have made this organism particularly useful for studying the principal mechanisms of development and identifying gene functions. Studying pathways and genetic interactions has provided the key to understanding many evolutionarily conserved developmental pathways, and it is now clear that of all invertebrate model organisms, the biology of *Drosophila* is by far the most closely related to that of humans.

The publication of almost the complete euchromatic part of the genome is a major achievement by a private-public partnership and will spark a new wave of interest in *Drosophila* as a model organism. The focus over the coming years will be on developing new ways to integrate the genomic data into the existing genetic experimental framework, and vice versa. Here we summarize some of the history of the *Drosophila* genome project, discuss aspects of the recently published complete sequence, and finally focus on how the genome sequence will influence future research on *Drosophila* (Fig. 1).

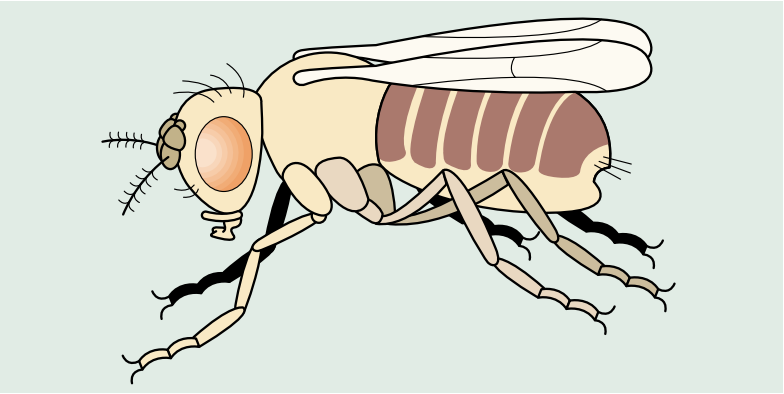

Figure 1 ***Drosophila melanogaster*. Direct sequencing and identification of the complete *Drosophila* genome will enhance its profile as a model organism and undoubtedly increase the pace of biological research.**

Historical landmarks

Efforts to map, sequence and annotate the *Drosophila* genome were born from the understanding of several research groups that its availability would greatly facilitate the molecular characterization of developmental genes and the analysis of genome organization. Analysis of the *Drosophila* genome is a long-standing tradition and in many instances studies of *Drosophila* have led to major conceptual and technical breakthroughs[1], such as the pioneering of physical mapping and saturation screens. As early as 1913, Sturtevant constructed a physical map of the *Drosophila* genome showing for the first time that genes were arranged in a linear order, and *Drosophila* has remained the organism with the most precise physical map since then. Saturation screens for mutations associated with embryonic defects, and the subsequent cloning of the corresponding genes revealed components of almost all known signalling pathways. Furthermore, many methods of manipulating genomes, such as the generation of transgenic animals, the use of transposable elements for mutagenesis and detection of expression patterns, site-specific recombination to rearrange chromosomes, and two-component control systems for ectopic gene expression, have their origins in *Drosophila* studies.

About nine years ago, the EDGP (European *Drosophila* Genome Project) and the BDGP began to generate a genome-wide clone coverage, on the basis of cloning strategies used for cosmid, YAC, P1 and BAC. These clones, freely available to the research community, are a unique resource for the mapping and positional cloning of genes. The BDGP also initiated a project to recover a collection of full-length sequenced complementary DNAs and expressed-sequence tags (ESTs). To date, cDNAs and ESTs representing, respectively, 40% and 65% of all *Drosophila* genes have been identified[2]. To sequence the *Drosophila* genome, both EDGP and BDGP followed a 'clone-by-clone' strategy and by 1999 approximately 20% of the euchromatic part of the genome had been deciphered. It was anticipated that the whole euchromatic sequence would be available by 2001/2.

However, the project received a boost in late 1998 when Celera decided to use the *Drosophila* genome as a proof of principle for the 'shotgun' sequencing of large eukaryotic genomes, which relies on breaking a genome into small random pieces that are then sequenced and reassembled by computational methods. This strategy, which had been successfully used on small prokaryotic genomes, met with mixed responses from the scientific community when Celera claimed that it could be used to sequence larger genomes. In less than a year, however, Celera had produced a whole-genome sequence with 6.5-times coverage. With the help of the BDGP, they assembled the *Drosophila* genome sequence, using information obtained by clone-based and shotgun sequencing, as well as STS generated from BACs[3]. The BDGP's overall contribution was 25 million bases (Mb) of the finished sequence as well as additional shotgun sequences of mapped BACs, to which the EDGP added 3 Mb of the X-chromosome sequence. The hybrid strategy of shotgun sequencing and physical mapping has proved successful in reducing the time and expense of whole-genome sequencing and may set an example for the sequencing of other complex genomes.

The *Drosophila* genome

The *Drosophila* genome has an estimated size of 180Mb, about 120 Mb of which is present as gene-rich euchromatin[4]. The joint project has successfully assembled 117 Mb in scaffolds and mapped them to chromosomes. More than 95% of the total sequence now resides in scaffolds between 100 thousand bases and 1 Mb in length, and 65% in scaffolds exceeding 10 Mb in size. However, at least 1301 gaps among mapped scaffolds remain to be filled. In addition, almost all of the heterochromatic sequence, which consists mainly of repetitive sequences that cannot be stably cloned, remains inaccessible to current methods.

Gene-prediction algorithms and searches of EST and protein databases predict that the *Drosophila* genome contains 13,601 protein-coding genes. This prediction, which is close to the previously estimate of 12,000 (ref. 5), relies on the assumption that the 60 Mb of heterochromatin does not contain higher-than-expected numbers of protein-coding genes. Perhaps surprisingly, the number of *Drosophila* genes is significantly lower than the 19,405 coding regions predicted for the smaller *Caenorhabditis elegans* genome. In contrast to *Drosophila, C. elegans* has a high number of local gene duplications that may account for much of this difference.

Rubin *et al.*[6] compared the sizes of the non-redundant, or 'core' proteomes encoded by yeast, *C. elegans* and *Drosophila*. The yeast genome seems to contain some 4,300 'core proteome' genes, and the *Drosophila* and *C. elegans* genomes 8000 and 9500, respectively. Interestingly, the core proteome of *Drosophila* is only twice as large as that of yeast and is smaller than in *C. elegans*. As differences in morphological and behavioural complexity are not correlated with gene numbers, determining why the *C. elegans* core proteome is more complex than that of *Drosophila* will be a challenge for the future.

The *Drosophila* sequence has been extensively annotated[6] and a wealth of information is now available to help to answer questions about genomic organization, development, cell biology, neurobiology, behaviour and evolution. Interestingly, comparisons with human sequences suggest that the *Drosophila* coding genome is more similar to humans than those of yeast and *C. elegans* are. This is illustrated by sequence searches with 289 human cancer-related genes, of which 61% have orthologues in *Drosophila*, in particular genes for MEN (multiple endocrine neoplasia), ATM (ataxia telangiectasia) and a p53-like protein. Analysis of the *Drosophila* genome also revealed several new homologues of signalling factors that are critical to developmental pathways, including two TGF-β proteins and three Wnt-family members that were not characterized by previous molecular or genetic analyses. Analysis of new genes will undoubtedly be a priority for many research groups. It is noteworthy that about 30% of the *Drosophila* proteome is not similar to any known genes. Although we must wait for more genomes to be completely sequenced, this finding will provide new insights into the evolution of animals.

Future directions for *Drosophila*

The availability of the full *Drosophila* genome will immediately affect the way experiments are conducted in the field. It will also stimulate new approaches and the development of new technologies (see Fig. 2).

The sequence information will save a huge amount of time with regard to the mapping of mutations and cloning of genes. The genomic information will help to guide and speed genetic analyses. For example, the knowledge that gene *X* is duplicated may explain why its mutant phenotype is weaker than expected. This genomic information would then allow the design of specific genetic screens to disrupt gene *X* and its homologue(s). Information obtained from the genome will also lead to the development of new approaches in functional genomics. In particular, the identification of all transcriptional units will allow the construction of a complete *Drosophila* microarray, as is already available in other organisms. Whole-genome transcriptional profiling will facilitate studies of global gene regulation and provide information on tissue- and cell-type-specific gene expression.

The analysis of databases for families of proteins with similar motifs will allow, in combination with gene-interference methods such as RNAi, the systematic functional analysis of entire gene families, by analysing the phenotypes, either singly or in combination, of all genes that contain a common protein domain (such as kinase, phosphatase, PDZ or SH2). *Drosophila* will also be useful in establishing the functions of mammalian genes with *Drosophila* orthologues. As characteristic embryonic or adult phenotypes are associated with most of the main signalling pathways, specific genes can be functionally linked to a pathway on the basis of their mutant phenotypes. For example, *Drosophila presinilin* mutants have a neurogenic phenotype reminiscent of loss of Notch activity. This suggests that human Presinilin, which is associated with Alzheimer's disease, also functions in the mammalian Notch pathway. Thus, many of the newly discovered *Drosophila* orthologues of human disease genes can be used to identify the pathways in which their encoded products are involved. In addition, genetic screens commonly used in *Drosophila* to observe genetic interactions may lead to the identification of further candidate protein factors in human diseases.

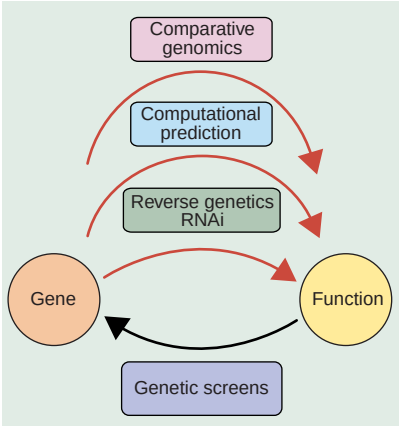

Figure 2 **Genome sequencing 'reverses genetics'.Classical genetics involved generating mutant phenotypes and identifying their genes. In this model the sequence of events is reversed: the functions of thousands of newly identified genes are yet to be elucidated.**

As more genome sequences become available, comparative genomics will be an increasingly useful approach for pinpointing different and common genes across species. For this kind of analysis, the sequence of a related species, such as *Drosophila virilis*, would be a valuable tool. Genome comparisons between different organisms will be informative on several levels, and information on genomic sequence and organization will be useful for exploring gene functions. Furthermore, the absence of gene families or pathway components from an organism's genome could prove informative about the necessity of its function in another organism.

The completion of the *Drosophila* genomic sequence is a major milestone, both for genomics, as it vindicates a new strategy for sequencing large eukaryotic genomes, and for *Drosophila*, as a model system to understand biological functions. As we enter the *Drosophila* postgenomic age, many old and new questions can now be tackled using this wonderful resource. □

Michael Boutros and Norbert Perrimon are in the Department of Genetics, Harvard Medical School, 200 Longwood, Boston, Massachusetts 02115, USA
Norbert Perrimon is also at the Howard Hughes Medical Institute, Harvard Medical School, 200 Longwood, Boston, Massachusetts 02115, USA
e-mail: perrimon@rascal.med.harvard.edu

1. Rubin, G. M. and Lewis, E. B. *Science* **287**, 2216–2218 (2000).
2. Rubin, G. M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E. *et al. Science* **287**, 2222–2224 (2000).
3. Myers, E. W. *et al. Science* **287**, 2196–2204 (2000).
4. Adams, M. D. *et al. Science* **287**, 2185–2915 (2000).
5. Miklos, G. L. G. and Rubin, G. M. *Cell* **86**, 521–529 (1996).
6. Rubin, G. M. *et al. Science* **287**, 2204–2215 (2000).