

Fly meets shotgun: shotgun wins

Daniel L. Hartl

Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.
e-mail: dhartl@oeb.harvard.edu

The sequencing group at Celera Genomics Corporation has cooperated with the European, Canadian and American genome projects to demonstrate that a genomic sequence as large as that of *Drosophila melanogaster* euchromatin (120 Mb) can be obtained by assembling the sequence of random ('shotgun') clones without the prerequisite for a clone-based physical map.

Like Lewis Carroll's White Queen, Craig Venter (President of Celera) sometimes seems to be able to believe as many as six impossible things before breakfast. Among his peer genome sequencers, the initial reaction to Venter's beliefs is usually harrumphing and pooh-poohing. But Venter's ideas are audacious rather than impossible, and he has established a record of delivering the goods on time.

First, he claimed that useful information could be obtained more quickly and cheaply by sequencing cDNAs than by sequencing a genome, an idea from which emerged his immensely important EST data¹. Then he said his group at The Institute for Genomic Research could sequence bacterial genomes by shotgun sequencing, in which the DNA is randomly cloned and sequenced and the overlaps used for assembly². (The alternative is clone-based sequencing, in which a significant amount of preliminary work is necessary to isolate individual clones and to order them according to the position in the genome from which the inserted DNA derives.) Only two years ago he claimed that his group at Celera Genomics Corporation could shotgun sequence the *Drosophila* genome, which, at 180 Mb, is the largest attempted so far. Four papers in the latest issue of *Science* attest to his latest success³⁻⁷.

The ends at the beginnings

The strategy of shotgun sequencing complex genomes requires sequence from each end of cloned inserts of various sizes⁸. This strategy facilitates assembly because each end-sequence is at a known distance and orientation from its partner. Included are some clones with large inserts that bridge across repeated sequences which would otherwise thwart assembly. In applying the method to *Drosophila*, the Celera group assembled sequences from the ends of three types of clones with insert lengths averaging 2 kb, 10 kb or 130 kb. Most of the sequence information is derived from

the 2-kb inserts, and much of the assembly information comes from the larger ones. The 10-kb inserts bridge across the length of most retrotransposons, which constitute the majority of repeated sequences in *Drosophila* euchromatin, whereas the 130-kb inserts provide long-range linking

additional sequence (2.5 Mb). The two sets of scaffolds, combined, yielded 117.3 Mb of euchromatin sequence. An extra 591 scaffolds totalling 2.4 Mb could not be assigned positions in the genomic sequence and may represent islands of complex sequence embedded in the 60 Mb of largely unclonable, highly repetitive DNA in centromeric heterochromatin. Nearly all of the 2,783 *Drosophila* genes previously sequenced are represented in the 247 scaffolds.

A significant amount of finishing work has yet to be carried out, because 1,798 gaps separate the 1,804 contigs (the average contig size is 65 kb) in the scaffolds. The average gap is estimated at 1,977 bp, yielding a total euchromatic genome size of about 120 Mb. These gaps are presently being filled through further efforts by the individual *Drosophila* genome projects.

The Annotation Jamboree

Accompanying the new genomic sequence is the accumulated wisdom gleaned from 90 years of investigations by *Drosophila* geneticists and other biologists (Fig. 1). To make use of the published and unpublished *Drosophila* lore, Celera hosted a two-week international 'annotation jamboree' of more than 40 scientists, primarily *Drosophila* aficionados, to define and classify the genes according to function^{3,5}. What did they find? *Drosophila* has 13,601 predicted genes, fewer than the number predicted in the nematode *Caenorhabditis elegans*⁹ (18,424), but in remarkable agreement with that predicted by classical studies of spontaneous mutation¹⁰. The giant polytene chromosomes in the salivary glands (Fig. 2) feature about 5,000 bands¹¹, so there are, on average, 2-3 genes per band. More than 5,500 of the genes are duplicated, accounting in part for the frequent polytene-band 'doublets' (adjacent, nearly identical bands) to which Calvin Bridges first drew attention¹². The tran-

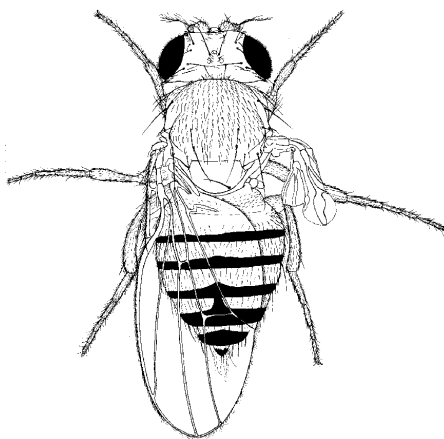


Fig. 1 Mosaic fly with one vestigial wing found by Calvin Bridges on 6 August 1923, and drawn by Edith M. Wallace. (Image reproduced with permission from D. Hartl.)

information. Altogether, more than 3 million sequence reads were analysed and assembled.

Scaffolds as gappy contigs

Assembly results in a set of scaffolds⁴. Each scaffold is a 'gappy contig', in which each sequence gap has a known location and estimated size because it is bridged by cloned inserts of known dimension whose ends are included in the scaffold. Assembly of the whole-genome shotgun sequences resulted in 50 scaffolds spanning 114.8 Mb that could be assigned positions on chromosomes. Inclusion of the 29 Mb of sequence (primarily from P1 clones and cosmids) previously obtained from the European, Canadian and American *Drosophila* genome projects added a large number of scaffolds (197), but not much

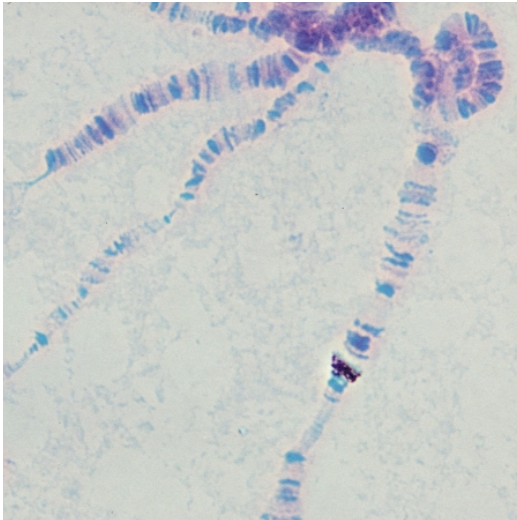


Fig. 2 *In situ* hybridization of a P1 clone with a site in a polytene salivary-gland chromosome. (Image reproduced with permission from D.L. Hartl and E.R. Lozovskaya.)

scripts average 3,058 bp and account for only one-third of the euchromatin.

Drosophila proteins can be assigned to 8,065 distinct families (which include the 5,500 duplicated genes) constituting the 'core proteome'⁵. About 5,000 distinct families in the core proteome are shared with *C. elegans*, and many protein families are highly elaborated. For example, *Drosophila* has about 300 protein kinases and about 80 protein phosphatases, 352 zinc-finger proteins of the C2H2 type, 199 trypsin-like serine proteases, 178 chymotrypsin-like (S1) proteases, 167 G-protein β -WD-40 repeat proteins, and 153 proteins with similarity to immunoglobulins and proteins of the major histocom-

patibility complex. Among the 700 transcription factors, 170 have previously been isolated and characterized. Some highly elaborated protein families are also found in *C. elegans*, but only three of the ten most highly elaborated families are in common between the two species.

Are physical maps obsolete?

Among 289 human genes that are mutated or show altered dosage or expression in human diseases, 61% have orthologues in flies. Among these are single-copy genes encoding previously unrecognized counterparts of proteins associated with cancer and neurological disorders, including p53, menin, frataxin and parkin. The analysis of the *Drosophila* proteome includes much more information of more specialized interest⁵. As in other organisms, about one-third of all *Drosophila* proteins have no significant similarity with any known proteins. What do these proteins do, and why do they apparently evolve so fast?

In their history of *Drosophila* genome landmarks, Rubin and Lewis⁷ lament that the initial success in cloning individual genes in *Drosophila* dampened enthusiasm for an organized genome project. I can attest to the truth of this from personal experience. When my laboratory first began assigning YAC and P1 clones to chromosomal locations by means of *in situ* hybridization¹³⁻¹⁵ (Fig. 2), influential members of the *Drosophila* community

would often ask publicly, "Why do we need a physical map when we already have the polytene chromosomes?" One answer was that mapped clones are immensely useful for isolating and studying individual genes and their expression. Another reality was that the technology for genomic sequencing available at the time absolutely required a clone-based physical map, and in 1992 we provided 19,200 P1 clones, many of them mapped, to the Rubin group as part of a collaborative effort. Many of these clones were widely distributed and proved to be useful to hundreds of *Drosophila* geneticists. Although clone-based physical maps still have many important uses in molecular genetics, Celera's success with the latest, fastest, cheapest high-throughput capillary sequencing technology may supersede the previously essential role of complete, clone-based physical maps in genomic sequencing. □

1. Fleischmann, R.D. *et al.* *Science* **269**, 496-512 (1995).
2. Adams, M.D. *et al.* *Nature* **377** (suppl.), 3-174 (1995).
3. Adams, M.D. *et al.* *Science* **287**, 2185-2195 (2000).
4. Myers, E.W. *et al.* *Science* **287**, 2196-2204 (2000).
5. Rubin, G.M. *et al.* *Science* **287**, 2204-2215 (2000).
6. Rubin, G.M. *et al.* *Science* **287**, 2216-2218 (2000).
7. Rubin, G.M. & Lewis, E.B. *Science* **287**, 2222-2224 (2000).
8. Weber, J.L. & Myers, E.W. *Genome Res.* **7**, 401-409 (1997).
9. *C. elegans* Sequencing Consortium. *Science* **282**, 2012-2018 (1998).
10. Muller, H.J. *Proc. Natl. Acad. Sci. USA* **14**, 714-726 (1928).
11. Sorsa, V. *Chromosome Maps of Drosophila* (CRC Press, Boca Raton, Florida, 1991).
12. Bridges, C.B. *J. Hered.* **29**, 11-13 (1938).
13. Garza, D. *et al.* *Science* **246**, 641-646 (1989).
14. Hartl, D.L., Nurminsky, D.I., Jones, R.W. & Lozovskaya, E.R. *Proc. Natl. Acad. Sci. USA* **91**, 6824-6829 (1994).
15. Hartl, D.L. & Lozovskaya, E.R. *The Drosophila Genome Map: A Practical Guide* (ed. Landes, R.G.) (Austin, Texas, 1995).

Oct-4, Scene 1: the drama of mouse development

Colin L. Stewart

Laboratory of Cancer and Developmental Biology, National Cancer Institute - ICRDC, Frederick, Maryland 21702, USA.
e-mail: stewartc@ncifcrf.gov

A study by Niwa *et al.* shows that changes in the levels of the transcription factor Oct-4 regulate the differentiation of embryonic stem cells along three different pathways. These observations suggest a possible mechanism by which Oct-4 determines the formation of the mouse pre-implantation embryo.

Compared with the frenetic development of other embryos, pre-implantation development of the mouse proceeds at relatively leisurely pace. This is a consequence of mammals being both viviparous (that is,

embryos develop inside the mother) and dependent on a placenta. Mammalian embryos first have to develop as 'pre-embryos', in which the tissue rudiments are established that enable the embryo to

form an effective union with the uterus. Pre-implantation development culminates in the formation of a blastocyst consisting of three tissues. A hollow epithelial sphere (the trophoblast), destined to form the