# *Drosophila* genome sequence

| ArticleInfo | | |
| --- | --- | --- |
| ArticleID | : | 3572 |
| ArticleDOI | : | 10.1186/gb-2000-1-1-reports038 |
| ArticleCitationID | : | reports038 |
| ArticleSequenceNumber | : | 63 |
| ArticleCategory | : | Paper report |
| ArticleFirstPage | : | 1 |
| ArticleLastPage | : | 4 |
| ArticleHistory | : | RegistrationDate : 2000–3–24<br>Received : 2000–3–24<br>OnlineDate : 2000–5–11 |
| ArticleCopyright | : | BioMed Central Ltd2000 |
| ArticleGrants | : | |
| ArticleContext | : | 130591111 |

Alan Shirras

## Abstract

The nucleotide sequence of the 120 megabase euchromatic region of the *Drosophila melanogaster* genome is reported.

# Significance and context

The completion of the *Drosophila melanogaster* genome sequence is a significant milestone for several reasons. First, it is the culmination of almost a century of genetic and molecular studies of this model organism; second, it has allowed the applicability of whole-genome sequencing to complex multicellular eukaryotic genomes to be tested; and third, it allows meaningful comparisons with the already sequenced *Caenorhabditis elegans* and *Saccharomyces cerevisiae* genomes to be made. Finally, the identification and annotation of *Drosophila* gene sequences will provide the basis for many future studies.

# Key results

Whole-genome sequencing was carried out by cloning size-selected, randomly sheared *Drosophila* genomic DNA into plasmid vectors and determining approximately 500 base pairs of sequence from either end. Overlapping stretches of sequence were then assembled into contiguous lengths. A crucial feature of this process was the use of the 'mate pairs' - stretches of sequence from either end of each clone - to minimize the problems of placing sequences containing repetitive DNA. The overall structure of the assembly was confirmed by linking the data to end-sequence information from bacterial artificial chromosomes (BACs) generated by the Berkeley *Drosophila* genome project (BDGP). This yielded 114.8 megabases (Mb) of sequence that could be unambiguously placed onto chromosomes. Clone-based sequence from the BDGP and the European *Drosophila* genome project (EDGP) allowed a further 1.4 Mb to be placed on chromosome arms. Roughly 3.8 Mb of sequence, probably representing islands of unique sequence within heterochromatin, could not be placed accurately on the map. By comparison with regions of high-quality sequence already determined by other methods, the whole-genome sequencing was found to be 99.99% accurate in non-repetitive regions. As a measure of the completeness of the sequence, 97.5% of sequenced *Drosophila* genes are found in the assembled sequence. Gene prediction by computational analysis followed by human curation has identified 13,601 genes. Of these, 23% do not match sequences from other organisms or from *Drosophila* expressed

sequence tags (ESTs) and are therefore potentially novel genes. Comparison of gene sequences with other species in general reveals a high degree of conservation although there are some exceptions; for example, several proteins involved in DNA repair are missing from *Drosophila*. A large number of transcription factors have been identified, suggesting complex networks of gene regulation, and solute transporters are also notable for their abundance and variety.

# Links

Coding content of the fly genome is summarized at the *Science* website. The CeleraScience website has *Drosophila* sequences for download as well as commentary on the *Drosophila* sequencing project. The *Drosophila melanogaster* genome sequence is also available from GenBank. The GadFly: Genome annotation database of *Drosophila* is available at Flybase.

# Conclusions

On the basis of their experience with *Drosophila* the authors point out that the BAC end-sequences and sequence-tagged sites (STS) content map were most cost-effective for long-distance sequence-based information and were necessary to link assemblies to chromosomal locations. A higher density of BAC end-sequences would have allowed larger sequence assemblies at lower shotgun coverage, and they recommend this for future projects.

# Reporter's comments

This sequence is of major benefit to *Drosophila* researchers among others and will open up many new avenues of research. Although considered "Release 1" by the authors, it is substantially complete and accurate. The whole-genome sequencing strategy has been shown to be applicable to a complex genome, and it seems likely that it will not be long before larger genomes, including the human, are completed in this fashion.

# Table of links

Science

Berkeley *Drosophila* genome project

European *Drosophila* genome project

Coding content of the fly genome

CeleraScience

*Drosophila melanogaster* genome

GadFly: Genome annotation database of *Drosophila*

# References

1.  Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al: The genome sequence of *Drosophila melanogaster*. Science. 2000, 287: 2185-2195. 0036-8075

This PDF file was created after publication.