The Drosophila Genome
Drosophila and developmental biology

the antennapedia mutant
Drosophila polytene chromosomes: a genome laid bare
The Genome Sequence of Drosophila melanogaster

Mark D. Adams,1* Susan E. Celniker,2 Robert A. Holt,1 Cheryl A. Evans,1 Jeannine D. Gocayne,1
Peter G. Amanatides,1 Steven E. Scherer,3 Peter W. Li,1 Roger A. Hoskins,2 Richard F. Galle,2 Reed A. George,2
Suzanna E. Lewis,4 Stephen Richards,2 Michael Ashburner,3 Scott N. Henderson,1 Granger G. Sutton,1
Jennifer R. Wortman,1 Mark D. Yandell,1 Qing Zhang,1 Lin X. Chen,1 Rhonda C. Brandon,1 Yu-Hui C. Rogers,1
Robert G. Blazej,5 Mark Champe,2 Barret D. Pfeiffer,1 Kenneth H. Wan,2 Clare Doyle,2 Evan G. Baxter,2
Gregg Helt,6 Catherine R. Nelson,4 George L. Gabor Miklos,3 Josep F. Abril,8 Anna Agbayani,2 Hui-Jin An,1
Cynthia Andrews-Pfannkoch,1 Danita Baldwin,1 Richard M. Ballew,1 Anand Basu,1 James Baxendale,1
Leyla Bayraktaroglu,9 Ellen M. Beasley,1 Karen Y. Beeson,1 P. V. Benos,10 Benjamin P. Berman,2 Deepali Bhandari,1
Slava Bolshakov,5 Dana Borkova,12 Michael R. Botchan,13 John Bouck,2 Peter Brokstein,4 Phillippe Brottier,14
Kenneth C. Burtis,15 Dana A. Busam,1 Heather Butler,16 Edouard Cadieu,17 Angela Center,1 Ishwar Chandra,1
J. Michael Cherry,18 Simon Cawley,19 Carl Dahlke,1 Lionel B. Davenport,1 Peter Davies,1 Beatriz de Pablos,20
Arthur Delcher,1 Zhuming Deng,1 Anne Deslattes Mays,1 Ian Dew,1 Suzanne M. Dietz,1 Kristina Dodson,1
Lisa E. Doup,1 Michael Downes,21 Shannon Dungan-Rocha,3 Boris C. Dunkov,22 Patrick Dunn,1 Kenneth J. Durbin,3
Carlos C. Evangelista,1 Concepcion Ferraz,23 Steven Ferriera,1 Wolfgang Fleischmann,2 Carl Fosler,1
Andrei E. Gabrielian,1 Neha S. Garg,1 William M. Gelbart,2 Ken Glasser,1 Anna Glodek,1 Fangcheng Gong,1
J. Harley Gorrell,3 Zhiping Gu,1 Ping Guan,1 Michael Harris,1 Nomi L. Harris,2 Damon Harvey,4 Thomas J. Heiman,1
Judith R. Hernandez,3 Jarrett Houck,1 Damon Hostin,1 Kathryn A. Houston,2 Timothy J. Howland,1 Ling-Hui Wei,1
Chinyye Ibegwam,24 Mena Jalali,1 Francis Kalush,1 Gary H. Karpen,21 Zhaoxi Ke,1 James A. Kennison,24
Karen A. Ketchum,1 Bruce E. Kimmel,2 Chinnappa D. Kodira,1 Cheryl Kraft,1 Saul Kravitz,1 David Kulp,6
Zhongwu Lai,1 Paul Lasko,23 Yiding Lei,1 Alexander A. Levitsky,1 Jiayin Li,1 Zhenya Li,1 Yong Liang,1 Xiaoying Lin,26
Xiangjun Liu,1 Bettina Mattei,1 Tina C. McIntosh,1 Michael P. McLeod,3 Duncan McPherson,1 Gennady Merkulov,1
Natalia V. Milshina,1 Clark Mobarry,1 Joe Morris,6 Ali Moshrefi,2 Stephen M. Mount,27 Mee Moy,1 Brian Murphy,1
Lee Murphy,28 Donna M. Muzny,3 David L. Nelson,1 David R. Nelson,29 Keith A. Nelson,1 Katherine Nixon,30
Deborah R. Nusskern,1 Joanne M. Pacleb,2 Michael Palazzolo,2 Gjanje S. Pittman,1 Sue Pan,1 John Pollard,1
Vinita Puri,1 Martin G. Reese,24 Knut Reinert,1 Karin Remington,1 Robert D. Saunders,31 Frederick Scheeler,1
Hua Shen,6 Bixiang Christopher Shue,1 Inga Siden-Kiamos,11 Michael Simpson,1 Marian P. Skupski,1 Tom Smith,1
Eugene Spier,1 Allan C. Spradling,11 Mark Stapleton,2 Renee Strong,1 Eric Sun,1 Robert Svirska,32 Cyndee Tector,1
Russell Turner,1 Eli Venter,2 Aihui H. Wang,1 Xin Wang,1 Zhen-Yuan Wang,1 David A. Wassarman,33
George M. Weinstock,1 Jean Weissendenbach,14 Sherita M. Williams,1 Trevor Woodage,1 Kim C. Worley,3 David Wu,1
Song Yang,5 Q. Alison Yao,1 Jane Ye,1 Ru-Fang Yeh,19 Jyashree S. Zaveri,1 Ming Zhao,1 Guoyang Zhang,1 Qi Zhao,1
Liansheng Zheng,1 Xiangqun H. Zheng,1 Fei N. Zhong,1 Wenyun Zhong,1 Xiaojun Zhou,1 Xiaoping Zhu,1
Xiaohong Zhu,1 Hamilton O. Smith,1 Richard A. Gibbs,3 Eugene W. Myers,1 Gerald M. Rubin,34 J. Craig Venter1

The fly Drosophila melanogaster is one of the most intensively studied organisms in biology and serves as a model system for the investigation of many developmental and cellular processes common to higher eukaryotes, including humans. We have determined the nucleotide sequence of nearly all of the ; 120-megabase euchromatic portion of the Drosophila genome using a whole-genome shotgun sequencing strategy supported by extensive clone-based sequence and a high-quality bacterial artificial chromosome physical map. Efforts are under way to close the remaining gaps; however, the sequence is of sufficient accuracy and contiguity to be declared substantially complete and to support an initial analysis of genome structure and preliminary gene annotation and interpretation. The genome encodes ; 13,600 genes, somewhat fewer than the smaller Caenorhabditis elegans genome, but with comparable functional diversity.

www.sciencemag.org SCIENCE VOL 287 24 MARCH 2000
The *Drosophila* Genome

180 Mb

divided into

~120 Mb of **euchromatic** DNA

and

~60 Mb of **heterochromatic** DNA

**HETEROCHROMATIN:**
densely staining, poorly uncondensed in interphase
heavily condensed during mitosis
GENE-POOR (by genetics)
TRANSPOSON- and REPEAT-RICH (by molecular biology)

**CENTROMERIC FUNCTION** flanked by heterochromatin
Mitotic chromosomes of D. melanogaster, showing euchromatic regions, heterochromatic regions, and centromeres.
The *Drosophila* Genome

$3 \times 10^6$ reads
The *Drosophila* Genome

180 Mb

commonest repeats are ~8 kb
so use a series of libraries:

2 kb plasmids - bulk of sequencing

10 kb plasmids - linking clones that SHOULD span any local repeat

130 kb BACs - clones also used for mapping
by fluorescence in situ hybridisation to polytene chromosomes
and STS content
Fig. 3. Assembly status of the Drosophila genome. Each chromosome arm is depicted with information on content and assembly status: (A) transposable elements, (B) gene density, (C) scaffolds from the joint assembly, (D) scaffolds from the WGS-only assembly, (E) polytene chromosome divisions, and (F) clone-based tiling path. Gene density is plotted in 50-kb windows; the scale is from 0 to 30 genes per 50 kb. Gaps between scaffolds are represented by vertical bars in (C) and (D). Clones colored red in the tiling path have been completely sequenced; clones colored blue have been draft-sequenced. Gaps shown in the tiling path do not necessarily mean that a clone does not exist at that position, only that it has not been sequenced. Each chromosome arm is oriented left to right, such that the centromere is located at the right side of X, 2L, and 3L and the left side of 2R and 3R.
Differences between WGS and clone-by-clone sequencing of *Drosophila* Adh region

- Thousands of base pairs gaps between contigs
- Missing from WGS
- Extra sequence in WGS
- Transposable elements
- Graph of mismatches

Thousands of base pairs
predicted 13601 protein-coding genes with 14113 predicted proteins (from alternative splices present in EST data)

on average, 1 gene per 9 kb
ranges from 0 to 30 genes per 50 kb window

NO obvious clustering by chromosomal location (arms, centres...)

a link between gene density and GC content
GC content drops as heterochromatin is approached, as does gene density (an effect spread over ~1 Mb)

24.1 Mb of the genome is coding (20% of euchromatin, 13% of whole)

54.1 Mb of the genome is transcribed as mRNA (37.5% of euchromatin, 24.5% of whole)
The *Drosophila* Genome

- 56673 exons
  - ~4 exons per gene
- 41000 introns
  - from 40 bp to 70000 bp
- each gene is
  - ~ 3000 bp

**BIGGEST PROTEIN:**
- "kakapo"
  - 5201 amino acids

**SMALLEST PROTEIN:**
- "ribosomal protein L38"
  - 21 amino acids

**ALSO predicted ...**
- 292 tRNAs
- 26 snRNAs
- ribosomal RNA clusters are part of heterochromatin
Table 1. Numbers of distinct gene families versus numbers of predicted genes and their duplicated copies in *H. influenzae*, *S. cerevisiae*, *C. elegans*, and *D. melanogaster*.

<table>
<thead>
<tr>
<th></th>
<th><em>H. influenzae</em></th>
<th><em>S. cerevisiae</em></th>
<th><em>C. elegans</em></th>
<th><em>D. melanogaster</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no. of predicted genes</td>
<td>1709</td>
<td>6241</td>
<td>18424</td>
<td>13601</td>
</tr>
<tr>
<td>No. of genes duplicated</td>
<td>284</td>
<td>1858</td>
<td>8971</td>
<td>5536</td>
</tr>
<tr>
<td>Total no. of distinct families</td>
<td>1425</td>
<td>4383</td>
<td>9453</td>
<td>8065</td>
</tr>
</tbody>
</table>
Table 5. Proteins in D. melanogaster, C. elegans, and S. cerevisiae with more than one InterPro domain.

<table>
<thead>
<tr>
<th>InterPro domains per protein</th>
<th>D. melanogaster (number of proteins)</th>
<th>C. elegans (number of proteins)</th>
<th>S. cerevisiae (number of proteins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>920</td>
<td>1236</td>
<td>410</td>
</tr>
<tr>
<td>3</td>
<td>388</td>
<td>458</td>
<td>121</td>
</tr>
<tr>
<td>4</td>
<td>219</td>
<td>182</td>
<td>58</td>
</tr>
<tr>
<td>5</td>
<td>163</td>
<td>98</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>101</td>
<td>72</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>92</td>
<td>53</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>58</td>
<td>27</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>42</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>11–15</td>
<td>73</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>16–20</td>
<td>18</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>21–30</td>
<td>22</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>31–50</td>
<td>8</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>51–75</td>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
FlyBase is a comprehensive database for information on the genetics and molecular biology of *Drosophila*. It includes data from the Drosophila Genome Projects and data curated from the literature. FlyBase is a joint project with the Berkeley Drosophila Genome Project.

FlyBase is supported by grants from the U.S. National Institutes of Health and the British Medical Research Council. See also the Warranty, Disclaimer & Copyright Notice.

Send comments to us at flybase-help@morgan.harvard.edu

http://www.flybase.org/
March 4, 2003: Whole genome alignment of D. pseudoobscura and D. melanogaster is now available via the Vista Genome Browser from the Berkeley Genome Pipeline.

January 31, 2003: The Release 3.1 annotated genomic sequence is now available for querying, sequence similarity searching, browsing, and download. Try our new Apollo genome annotation browser.

About BDGP
Contact information, news archive, how to cite BDGP, and general overview

Searches
Fly BLAST
Sequence similarity search of Drosophila-specific data sets

GadFly
Search/browse the Drosophila genome annotations, by name, chromosomal position, molecular function or protein domain

FlyBase All Searches
FlyBase, BDGP, EDGP

Analysis Tools
Search sequence for patterns, splice sites, etc.

Projects
Genomic Sequencing
Sequence and annotation of the Drosophila genome

Expression Patterns
Systematic determination of patterns of gene expression in Drosophila embryogenesis by RNA in situ

cDNAs & ESTs
Drosophila Gene Collection, full-length cDNA sequencing and expressed sequence tags (ESTs)

Transposon Insertions
Insertion of single P transposable elements to disrupt and misexpress Drosophila genes

Comparative Genomics
Comparative analysis of genomic sequences in Drosophila and related species

SNPs
Single Nucleotide Polymorphism Map

BDGP Resources
Download
Sequence data sets and annotations in fasta or xml format by http or ftp

Materials
Request genomic or cDNA clones, library filters or fly stocks

Publications
Browse or download BDGP papers

Methods
BDGP laboratory protocols and vector maps

Software
Download BDGP software

The Berkeley Drosophila Genome Project (BDGP) is a consortium of the Drosophila Genome Center (funded by the National Human Genome Research Institute, National Cancer Institute, and the Department of Energy), and the Howard Hughes Medical Institute (through its support of work in the Gerald Rubin and Allan Spradling laboratories).

This page last updated on: 3/4/03

Please send comments or questions about the web site to bdgp@fruitfly.org

http://www.fruitfly.org/
The European Drosophila Genome Project is a consortium consisting of 10 labs and has the objective of sequencing the X chromosome of *Drosophila melanogaster*. Divisions 1 to 3 were sequenced during this project. Sequencing was initially based on the cosmid physical map, constructed by the European Drosophila Physical Mapping Project; and consequently on BAC map from three genomic libraries.

### What have we done?

So far, EDGP has submitted 119 sequences (finished and unfinished) or 6,448,051 bp to the public data libraries. 3.2 Mbp of non-redundant sequence has been analysed and annotated, resulting in 243 newly discovered genes plus 86 previously sequenced/studied ones.

The analysis is following the scheme presented [here](http://edgp.ebi.ac.uk/). For lists of genomic sequences and annotation, please visit the following:

- Progress of the sequencing
- Progress of the annotation

### More *Drosophila* links...

- FlyBase
- BDGP

---

The European *Drosophila* Genome Project is funded by a contract from the Biotechnology Programme of the European Union (contract no.: BIO4-CT96-0506). It has been funded since January 1997.

This page was last updated on February 14, 2000 by Takis Benos

If you have any questions/comments, please, e-mail us at edgp@ebi.ac.uk. We'd love to hear from you.

http://edgp.ebi.ac.uk/