Published in final edited form as:
*Science*. 2007 June 15; 316(5831): 1625–1628. doi:10.1126/science.1139816.

# Sequence Finishing and Mapping of *Drosophila melanogaster* Heterochromatin

**Roger A. Hoskins**[1,*], **Joseph W. Carlson**[1,*], **Cameron Kennedy**[1], **David Acevedo**[1], **Martha Evans-Holm**[1], **Erwin Frise**[1], **Kenneth H. Wan**[1], **Soo Park**[1], **Maria Mendez-Lago**[2], **Fabrizio Rossi**[3], **Alfredo Villasante**[2], **Patrizio Dimitri**[3], **Gary H. Karpen**[1,4], and **Susan E. Celniker**[1,†]

[1] Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[2] Centro de Biologia Molecular Severo Ochoa, CSIC-UAM, Cantoblanco 28049, Madrid, Spain

[3] Dipartimento di Genetica e Biologia Molecolare “Charles Darwin,” Universita “La Sapienza,” 00185 Roma, Italy

[4] Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA

## Abstract

Genome sequences for most metazoans and plants are incomplete because of the presence of repeated DNA in the heterochromatin. The heterochromatic regions of *Drosophila melanogaster* contain 20 million bases (Mb) of sequence amenable to mapping, sequence assembly, and finishing. We describe the generation of 15 Mb of finished or improved heterochromatic sequence with the use of available clone resources and assembly methods. We also constructed a bacterial artificial chromosome–based physical map that spans 13 Mb of the pericentromeric heterochromatin and a cytogenetic map that positions 11 Mb in specific chromosomal locations. We have approached a complete assembly and mapping of the nonsatellite component of *Drosophila* heterochromatin. The strategy we describe is also applicable to generating substantially more information about heterochromatin in other species, including humans.

Heterochromatin is a major component of metazoan and plant genomes (e.g., ~20% of the human genome) that regulates chromosome segregation, nuclear organization, and gene expression (1–4). A thorough description of the sequence and organization of heterochromatin is necessary for understanding the essential functions encoded within this region of the genome. However, difficulties in cloning, mapping, and assembling regions rich in repetitive elements have hindered the genomic analysis of heterochromatin (5–7). The fruit fly *Drosophila melanogaster* is a model for heterochromatin studies. About one-third of the genome is considered heterochromatic and is concentrated in the pericentromeric and telomeric regions of the chromosomes (X, 2, 3, 4, and Y) (5,8). The heterochromatin contains tandemly repeated simple sequences (including satellite DNAs) (9), middle repetitive elements [such as transposable elements (TEs) and ribosomal DNA], and some single-copy DNA (10).

†To whom correspondence should be addressed. celniker@fruitfly.org.
*These authors contributed equally to this work.

Supporting Online Material
www.sciencemag.org/cgi/content/full/316/5831/1625/DC1
Materials and Methods
SOM Text
Figs. S1 to S8
References

The whole-genome shotgun sequence (WGS3) was the foundation for finishing and mapping heterochromatic sequences and for elucidating the organization and composition of the nonsatellite DNA in *Drosophila* heterochromatin (5,6). WGS3 is an excellent assembly of the *Drosophila* euchromatic sequence, but it has lower contiguity and quality in the repeat-rich heterochromatin. We undertook a retrospective analysis of these WGS3 scaffolds (11). Moderately repetitive sequences, such as transposable elements, are well represented in WGS clones and sequence reads, but they tend to be assembled into shorter scaffolds with many gaps and low-quality regions because of the difficulty of accurately assigning data to a specific copy of a repeat. The typical WGS heterochromatic scaffold is smaller [for scaffolds mapped to an arm, N50 ranged from 4 to 35 kb (11)] than a typical WGS euchromatic scaffold (N50 = 13.9 Mb) (5). Relative to the euchromatic scaffolds, the WGS3 heterochromatic scaffolds have 5.8 times as many sequence gaps per Mb, as well as lower sequence quality.

To produce the Release 5 sequence, we identified a set of 10-kb genomic clones from a library representing 15× clone coverage by paired end reads (mate pairs) and used this set as templates to fill small gaps and improve low-quality regions (11). Higher-level sequence assembly into Mb-sized linked scaffolds used relationships determined from bacterial artificial chromosome (BAC)–based sequence tag site (STS) physical mapping (see below) and BAC end sequences. In addition to the WGS data, we incorporated data from 30 BACs (3.4 Mb; 15 BACs finished since Release 3) that were originally sequenced as part of the euchromatin sequencing effort (5,10).

Sequence finishing resulted in fewer gaps, longer scaffolds, and higher-quality sequence relative to WGS3 (fig. S1). About 15 Mb of this sequence has been finished or improved, and 50% of the sequence is now in scaffolds greater than 378 kb (N50). Table 1 summarizes the Release 5 sequence statistics by chromosome arm. Improved sequence was generated for 145 WGS3 scaffolds, and a set of 90 new scaffolds were produced by joining or filling 694 gaps of previously unknown size between WGS3 scaffolds. The relationships between the initial WGS scaffolds and the Release 5 scaffolds can be complex (Fig. 1 and figs. S2 to S7); for example, there were eight cases in which small scaffolds were used to fill gaps within larger scaffolds, and two scaffolds whose gaps interdigitated. As expected, the sequence consists largely of nests of fragmented TEs, and most remaining gaps are bounded by TEs or simple sequence repeats, including simple repeats not previously described (Fig. 2). The quality of the improved sequence was measured by calculating the estimated error rates within 10-kb sliding windows (overlapping by 5 kb) on the consensus sequences (11). For all but 11 of 1832 10-kb regions not overlapping one of the known TEs, the estimated error rate is less than 1 per 17,986 base pairs (bp), well below the accepted standard for finished genomic sequence of 1 error per 10,000 bp.

Concurrent with the sequence-finishing effort, we constructed an integrated physical and cytogenetic map to describe the overall structure of the pericentromeric heterochromatin. This map was essential for ordering, orienting, and linking WGS sequence scaffolds into larger BAC contigs and Release 5 scaffolds. Heterochromatic sequences at the centric ends of the Release 3 arm sequences were represented in BAC-based physical maps of the euchromatic and telomeric portions of the chromosomes (12,13), but most heterochromatic scaffolds had not been mapped in large-insert clones or localized to specific sites on the chromosomes.

BAC-based STS content mapping of WGS3 scaffolds, using 354 probes designed from genomic sequence and five BAC libraries (11), extended and linked many scaffolds into larger BAC contigs. The BAC map incorporates scaffolds spanning 13.4 Mb of the WGS3 assembly and links 14 WGS3 scaffolds to the Release 3 arm sequences (Table 2). In regions proximal to the arm assemblies, it links 130 WGS3 scaffolds into 25 multiscaffold BAC contigs and yields 21 single-scaffold BAC contigs (Table 2) (11). The largest BAC contig links 20 WGS3 scaffolds spanning 1.7 Mb.

We used fluorescence in situ hybridization (FISH) to map BAC contigs and sequence scaffolds to specific cytogenetic locations in mitotic chromosomes (11,14). The high repeat content of heterochromatin required the use of single-copy probes [P-element insertions (15,16) and cDNA clones (17,18)] that could be assigned to specific sequence scaffolds. We also used BAC probes that had sufficient single-copy sequences to provide unambiguous localizations (11) (fig. S8). The physical and cytogenetic mapping results and previously published data were used to produce an integrated map of pericentromeric heterochromatin (11). We present cytogenetic locations for 15 BAC contigs linking 80 scaffolds and an additional 14 scaffolds that were linked to chromosome arms; these localized scaffolds span 11.2 Mb of pericentromeric heterochromatin in the WGS3 assembly (Table 2). Currently unlocalized are 50 WGS3 scaffolds in 31 BAC contigs, as well as an additional 63 WGS3 scaffolds larger than 15 kb that are not represented in the BAC map. Four scaffolds larger than 15 kb and not represented in the BAC map were incorporated into Release 5 by sequence finishing (11).

Integration of the map and sequence-finishing information led us to define three classes of Release 5 heterochromatic scaffolds: (i) contiguous with the assembled euchromatic arms and extending them farther into pericentromeric heterochromatin (chromosome arm “h”); (ii) mapped to specific chromosome arms with partial information on order and orientation and concatenated into “arm” files (arm “Het”); and (iii) unmapped and concatenated into a single file (arm “U”). The improved, mapped Release 5 scaffolds are diagrammed relative to the chromosome arms in Fig. 3; see (11) for analysis of sequences and maps by chromosome.

We have demonstrated substantial progress toward our goal of assembling and mapping the components of heterochromatin that are not simple repeats, and have shown that heterochromatic regions containing single-copy genes and a high density of transposable elements can be assembled into high-quality, contiguous sequence. How can we generate an even more complete genomic understanding of *Drosophila* heterochromatin? The tiling path of overlapping BACs spanning the Release 5 sequence (11) provides templates for gap closure and scaffold extension in the regions that contain middle-repetitive elements and single-copy genes. Progress can also be made in localizing more sequences by performing FISH with additional cDNAs, BACs, and transposon insertions from other collections (19,20). Restriction fingerprints of tiling path BACs will also provide an independent benchmark to evaluate the accuracy of finished sequence assemblies (21). The apparent absence of BACs covering various remaining gaps likely reflects the presence of extensive simple sequence arrays, which are unlikely to be completely closed as the map and sequence are improved. New technologies will be required to determine the sequence and structure of these highly repetitive regions. However, an achievable goal using current technologies is to produce complete maps and sequence assemblies for the single-copy and middle-repetitive components of the heterochromatin, combined with cytological definition of the locations and structures of large blocks of tandemly repeated simple-sequence DNA.

Our results suggest that elucidating the organization and composition of heterochromatic regions in other organisms is a practical goal. However, our ability to substantially improve the sequence and maps required three critical components: (i) a high-quality WGS sequence assembly; (ii) a high-depth collection of precisely sized and aligned genomic clones for sequence finishing and gap closure; and (iii) physical and cytogenetic mapping to deduce relationships between WGS scaffolds. The STS content-mapping experiments benefited greatly from the availability of large-insert BAC libraries produced by fragmenting genomic DNA with three different restriction enzymes and with physical shearing. Analysis of heterochromatin in other genomes would also benefit from improved algorithms that can successfully and accurately assemble sequence of regions rich in repeated DNA.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## References and Notes

1. Dernburg AF, et al. Cell 1996;85:745. [PubMed: 8646782]
2. Karpen GH, Le MH, Le H. Science 1996;273:118. [PubMed: 8658180]
3. Talbert PB, Henikoff S. Nat Rev Genet 2006;7:793. [PubMed: 16983375]
4. Wallrath LL. Curr Opin Genet Dev 1998;8:147. [PubMed: 9610404]
5. Celniker SE, et al. Genome Biol 2002;3:RESEARCH0079. [PubMed: 12537568]
6. Hoskins RA, et al. Genome Biol 2002;3:RESEARCH0085. [PubMed: 12537574]
7. Eichler EE, Clark RA, She X. Nat Rev Genet 2004;5:345. [PubMed: 15143317]
8. John, B.; Miklos, GLG. The Eukaryote Genome in Development and Evolution. Allen & Unwin; London: 1988. p. 416
9. Lohe AR, Brutlag DL. Proc Natl Acad Sci USA 1986;83:696. [PubMed: 3080746]
10. Adams MD, et al. Science 2000;287:2185. [PubMed: 10731132]
11. See supporting material on *Science* Online.
12. Hoskins RA, et al. Science 2000;287:2271. [PubMed: 10731150]
13. Abad JP, et al. Mol Biol Evol 2004;21:1613. [PubMed: 15163766]
14. Gatti M, Pimpinelli S. Annu Rev Genet 1992;26:239. [PubMed: 1482113]
15. Konev AY, et al. Genetics 2003;165:2039. [PubMed: 14704184]
16. Yan CM, Dobie KW, Le HD, Konev AY, Karpen GH. Genetics 2002;161:217. [PubMed: 12019236]
17. The *Drosophila* Gene Collection. (www.fruitfly.org/EST/index.shtml)
18. Stapleton M, et al. Genome Res 2002;12:1294. [PubMed: 12176937]
19. Bellen HJ, et al. Genetics 2004;167:761. [PubMed: 15238527]
20. Thibault ST, et al. Nat Genet 2004;36:283. [PubMed: 14981521]
21. Marra MA, et al. Genome Res 1997;7:1072. [PubMed: 9371743]
22. Gatti M, Bonaccorsi S, Pimpinelli S. Methods Cell Biol 1994;44:371. [PubMed: 7707964]
23. Celniker, SE., et al. BDGP: Release 5 Genomic Sequence Download. Mar. 2006 (www.bdgp.org/sequence/release5genomic.shtml)
24. We thank Celera Genomics Inc. for the 10-kb genomic clones that we used as sequencing templates, A. B. de Carvalho for discussions of Y chromosome sequences, and R. Svirskas, A. M. Ryles, E. Kym, R. Chetty, and S. Galle for technical assistance. Supported by NIH grant R01 HG00747 (G.H.K.); BAC-based sequencing was supported by NIH grant P50-HG00750 (G. M. Rubin) and U.S. Department of Energy contract DE-AC0376SF00098 (S.E.C.).

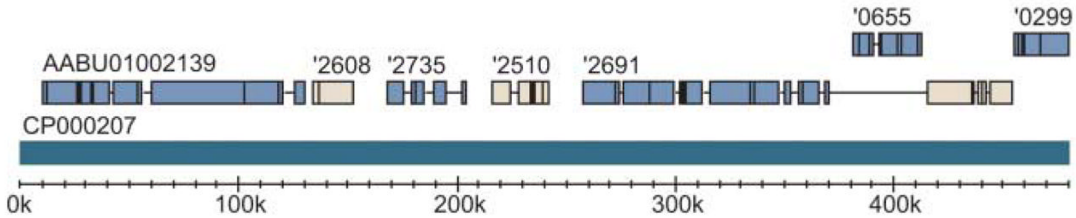


**Fig. 1.**
Comparison of WGS scaffolds to the corresponding Release 5 scaffold. WGS scaffolds (gray, same orientation; tan, opposite orientation) are diagrammed above the Release 5 scaffold (blue). Sequence gaps (thin horizontal lines) in WGS scaffolds are indicated.

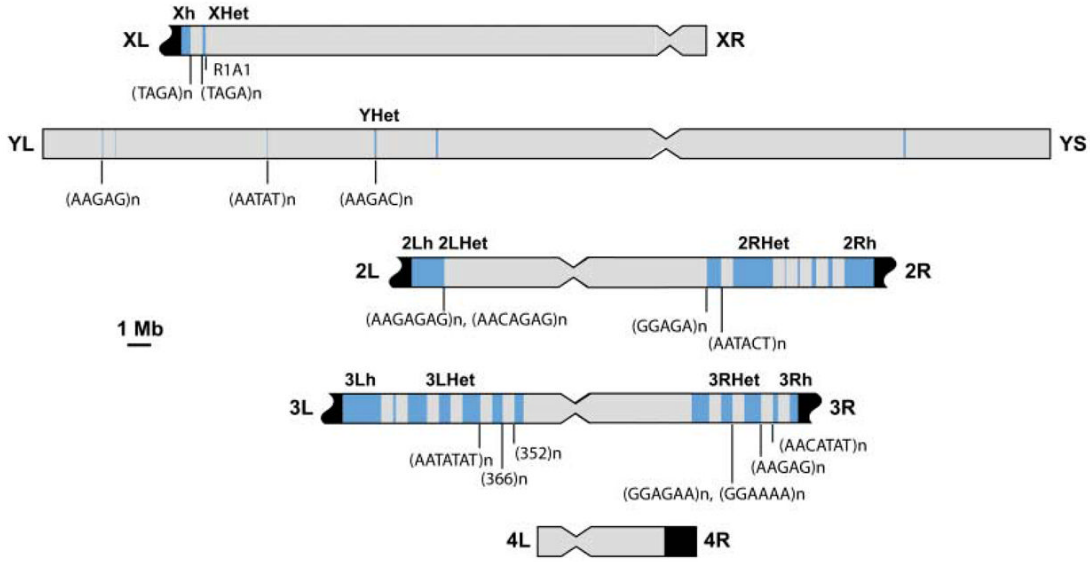


**Fig. 2.**

Sequenced regions of *D. melanogaster* pericentromeric heterochromatin. The heterochromatin extends proximally from the euchromatin (black) and includes sequenced and assembled regions (aqua) and unsequenced regions (gray). The actual gap sizes between sequence scaffolds are unknown and are presented with an arbitrary 0.5-Mb separation. Finished or improved scaffolds, which end in known or novel simple repeats, are shown with the terminal repeat sequence indicated. The scaffold CP000217, originally identified as part of 2RHet but subsequently mapped to 3LHet, is shown here at its updated location (see text).

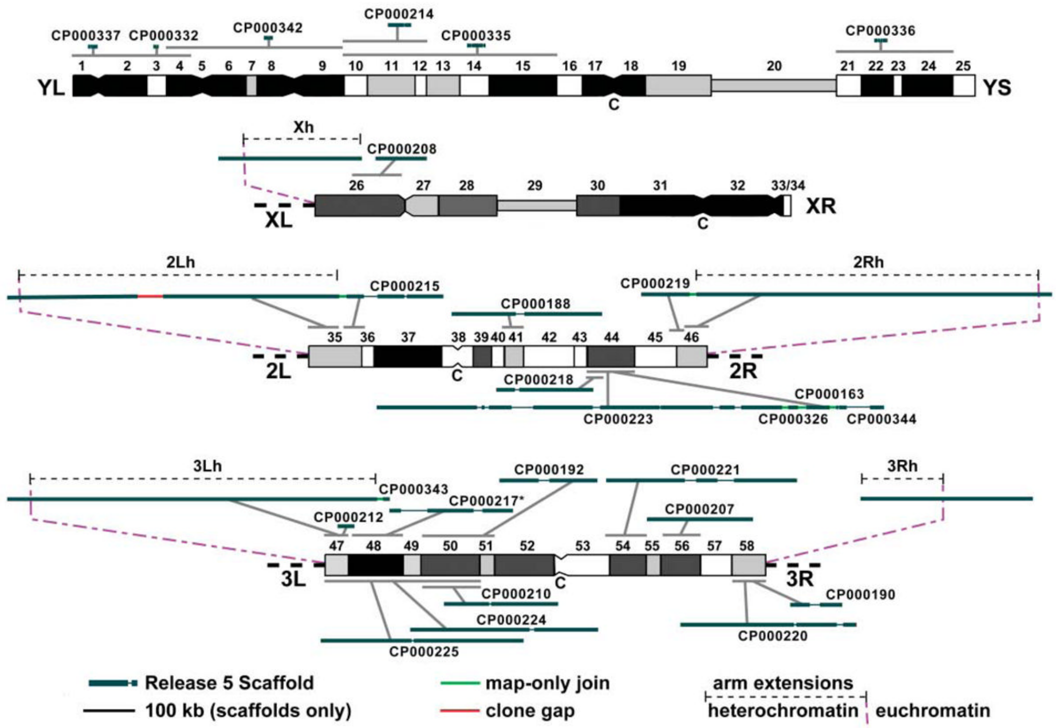


**Fig. 3.**

Integrated map of *D. melanogaster* pericentromeric heterochromatin. The cytogenetic reference map of the heterochromatic regions of the chromosomes with numbered divisions (h1 to h58) and centromeres (C) is shown (22). The fourth chromosome (h58 to h61) is not shown. Release 5 sequence scaffolds are indicated at their cytogenetic map locations, and Het scaffolds are labeled with their GenBank accession numbers. Scaffolds (13.9 Mb in total; see scale bar) and the heterochromatin (100 Mb in total) are represented at different scales. Sequence contigs (thick bars) and sequence gaps (thin bars) within scaffolds are shown. Some sequence gaps are too small to be represented at this scale. A clone gap in the 2Lh sequence is indicated. Joins between Release 5 scaffolds present in the BAC map assembly but not yet incorporated in the sequence assembly are shown. Cytogenetic locations are indicated by lines connecting scaffolds to cytogenetic ranges. The heterochromatin-euchromatin boundaries within the sequence of the chromosome arms, based on BAC FISH (6), are indicated by dashed magenta lines. The orientations of Het scaffolds are not necessarily known (11,23). CP000217, originally identified as part of 2RHet but subsequently mapped to 3LHet, is shown here at its updated location; CP000206, originally identified as part of 3RHet but subsequently removed to the unlocalized scaffolds, is not shown (11).

## Table 1

Status of Release 5. Sequence statistics for the chromosomes are divided into regions contiguous with the euchromatic arm sequences (e.g., Xh) and regions mapped cytologically to those chromosome arms but not currently connected (e.g., XHet). Bac-Based Rel. 5 refers to the amount of heterochromatin finished in BACs. N50 is the contig length such that 50% of all base pairs are contained in contigs of this length or larger.

| Region | Size (bp) | BAC-Based Rel. 5 | Rel. 5 without N's | N50 | Sized gaps | Total gap size | Unsized gaps |
|---|---|---|---|---|---|---|---|
| Xh | 392,502 | 312,439 | 392,502 | 392,502 | 0 | 0 | 0 |
| XHet | 204,112 | — | 204,112 | 204,112 | 0 | 0 | 0 |
| 2Lh | 1,010,570 | 1,010,470 | 1,010,470 | 591,203 | 0 | 0 | 1 |
| 2LHet | 368,872 | — | 297,872 | 99,162 | 2 | 71,000 | 0 |
| 2Rh | 1,285,689 | 973,874 | 1,285,689 | 1,285,689 | 0 | 0 | 0 |
| 2RHet* | 3,288,761 | — | 2,721,941 | 244,298 | 17 | 566,020 | 8 |
| 3Lh | 1,587,982 | 1,020,114 | 1,587,982 | 1,587,982 | 0 | 0 | 0 |
| 3LHet* | 2,555,491 | — | 2,416,308 | 366,456 | 12 | 138,483 | 7 |
| 3Rh | 378,656 | 378,656 | 378,656 | 378,656 | 0 | 0 | 0 |
| 3RHet | 2,517,507 | — | 2,264,306 | 252,624 | 10 | 252,801 | 4 |
| YHet | 347,038 | — | 242,806 | 9,129 | 30 | 101,632 | 26 |
| Unmapped modified | 2,419,890 | — | 2,222,443 | 73,591 | 15 | 194,247 | 32 |
| Total for modified sequence | 16,357,070 | — | 15,025,087 | 378,616 | 86 | 1,324,183 | 78 |
| Unmapped unmodified | 7,629,047 | — | 6,145,805 | 2,521 | 439 | 1,239,942 | 2,433 |
| Total | 23,986,117 | 3,383,114 | 21,170,892 | — | 525 | 2,564,125 | 2,511 |

* Statistics reflect the sequence distributed as Release 5 of the genome and do not account for the scaffold CP000217 moved from 2RHet to 3LHet and the scaffold CP000206 moved from 3RHet to ArmU subsequent to the release.

## Table 2

Summary of the integrated physical and cytogenetic map assembly. N/A, not applicable.

| Chromosome arm | STSs in BAC map | Failed STSs in mapped contigs | WGS3 scaffolds linked to chr. arm | WGS3 scaffolds in Het contigs | Het contigs | Sum of WGS3 lengths in mapped scaffolds (kb) |
|---|---|---|---|---|---|---|
| XL | 16 | 0 | 3 | 2 | 1 | 498 |
| 2L | 28 | 1 | 2 | 5 | 1 | 1,018 |
| 2R | 91 | 1 | 5 | 29 | 3 | 3,517 |
| 3L | 91 | 1 | 3 | 24 | 6 | 4,039 |
| 3R | 51 | 0 | 0 | 20 | 4 | 2,101 |
| 4R | 8 | 2 | 1 | 0 | 0 | 65 |
| Y | 1 | 4 | N/A | 0 | 0 | 0 |
| Subtotal | | | | | | |
| (Localized) | 286 | 9 | 14 | 80 | 15 | 11,238 |
| U | 68 | N/A | N/A | 50 | 31 | 2,177 |
| Total | 354 | 9 | 14 | 130 | 46 | 13,415 |