Research

Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly

Roger A Hoskins*†, Christopher D Smith‡†, Joseph W Carlson*, A Bernardo Carvalho§, Aaron Halpern¶, Joshua S Kaminker‡, Cameron Kennedy#, Chris J Mungall††, Beth A Sullivan#, Granger G Sutton¶, Jiro C Yasuhara**, Barbara T Wakimoto**, Eugene W Myers¶, Susan E Celniker*, Gerald M Rubin*‡††† and Gary H Karpen#

Addresses: *Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ‡Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. §Departamento de Genética, Universidade Federal do Rio de Janeiro, CEP 21944-970, Rio de Janeiro, Brazil. ¶Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. #Molecular and Cell Biology Laboratory, Salk Institute, La Jolla, CA 92037, USA. ††Howard Hughes Medical Institute, University of California, Berkeley, CA 94720, USA. **Department of Zoology, University of Washington, Seattle, WA 98195, USA. †These authors contributed equally to this work.

Correspondence: Roger A Hoskins. E-mail: rhoskins@lbl.gov

Published: 31 December 2002

Genome ***Biology*** 2002, **3(12)**:research0085.1–0085.16

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0085

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 30 October 2002
Revised: 28 November 2002
Accepted: 5 December 2002

Abstract

Background: Most eukaryotic genomes include a substantial repeat-rich fraction termed heterochromatin, which is concentrated in centric and telomeric regions. The repetitive nature of heterochromatic sequence makes it difficult to assemble and analyze. To better understand the heterochromatic component of the *Drosophila melanogaster* genome, we characterized and annotated portions of a whole-genome shotgun sequence assembly.

Results: WGS3, an improved whole-genome shotgun assembly, includes 20.7 Mb of draft-quality sequence not represented in the Release 3 sequence spanning the euchromatin. We annotated this sequence using the methods employed in the re-annotation of the Release 3 euchromatic sequence. This analysis predicted 297 protein-coding genes and six non-protein-coding genes, including known heterochromatic genes, and regions of similarity to known transposable elements. Bacterial artificial chromosome (BAC)-based fluorescence *in situ* hybridization analysis was used to correlate the genomic sequence with the cytogenetic map in order to refine the genomic definition of the centric heterochromatin; on the basis of our cytological definition, the annotated Release 3 euchromatic sequence extends into the centric heterochromatin on each chromosome arm.

Conclusions: Whole-genome shotgun assembly produced a reliable draft-quality sequence of a significant part of the *Drosophila* heterochromatin. Annotation of this sequence defined the intron-exon structures of 30 known protein-coding genes and 267 protein-coding gene models. The cytogenetic mapping suggests that an additional 150 predicted genes are located in heterochromatin at the base of the Release 3 euchromatic sequence. Our analysis suggests strategies for improving the sequence and annotation of the heterochromatic portions of the *Drosophila* and other complex genomes.

Background

Heterochromatin was first distinguished from euchromatin cytologically, on the basis of differential staining properties [1]. Molecular and genetic properties that further distinguish heterochromatin from euchromatin include DNA sequence composition, replication timing, condensation throughout the cell cycle, and the ability to silence gene expression [2-4]. In addition to genes required for viability and fertility [5], heterochromatin contains essential *cis*-acting chromosome inheritance loci, including elements required for centromere function [6], meiotic pairing [7-9], and sister chromatid cohesion [10,11]. A significant fraction of the fly and human genomes are heterochromatic, yet our current understanding of the sequence and organization of heterochromatin is very limited. Heterochromatin is concentrated in megabase-sized tracts in the centric and subtelomeric regions of the chromosomes. It contains tandemly repeated short sequences (satellite DNAs), middle repetitive elements (for example, transposable elements), and some single-copy sequences [4]. Progress has been made in the analysis of the non-satellite component of *Drosophila*, *Arabidopsis* and human heterochromatin [12-20]. Less progress has been made in analysis of satellite sequences, although recent studies have revealed the structure and composition of centromeric satellites [21,22].

Heterochromatin accounts for an estimated 59 megabases (Mb) of the 176-Mb genome of the *Drosophila melanogaster* female, and the 41-Mb male Y chromosome is entirely heterochromatic (Figure 1). Polytene chromosomes, which are so valuable in mapping euchromatic genes in *Drosophila*, provide minimal resolution in the heterochromatin. The heterochromatin is either severely under-represented or poorly banded in polytene chromosomes, and aggregates into an unbanded structure known as the chromocenter [23,24]. *Drosophila* heterochromatin is best defined by the cytogenetic map of neuroblast mitotic chromosomes, which resolves 61 heterochromatic bands (h1-h61) by differential staining properties [5,25]. The locations of repeated sequences within the heterochromatin have been studied using *in situ* hybridization of probes derived from known repeated elements to mitotic chromosomes [26,27]. This, together with molecular analysis of minichromosomes, has elucidated the general organization and composition of heterochromatin. Satellite blocks of 20 kilobases (kb) to 1 Mb are interrupted by 'islands' of 5-50 kb of complex sequences that contain a high density of transposable elements [22,28-30].

The transition between heterochromatin and euchromatin appears to be gradual rather than abrupt. For example, a hallmark of heterochromatin is a high density of transposable elements, and the density of these elements in the genomic sequence increases continuously toward the centric ends of the euchromatic portions of the chromosome arms [12,31,32]. This trend continues in the centric heterochromatin, which contains large blocks of specific types of middle-repetitive sequences [22,27].

Heterochromatic genes have been defined by mutations that affect viability or fertility [33]. Genetic screens, reviewed in [5], have identified 14 vital loci in the heterochromatin of chromosome 2 [34,35] and 12 vital loci in the heterochromatin of chromosome 3 [36]. Although no vital loci have been identified in the proximal heterochromatin of the X chromosome [37], several identified loci map near the boundary between the centric heterochromatin and the euchromatin (see [38]). There are six Y-linked loci required for male fertility ([39], reviewed in [5]). Thus, there are at least 32 identified genetic loci required for viability or fertility in the centric heterochromatin. This is likely to be an underestimate, because saturating genetic screens have not been done for all of the heterochromatin.

Molecularly characterized *Drosophila* heterochromatic genes encode diverse proteins and functions. Examples include *light* (post-Golgi protein trafficking), *concertina*

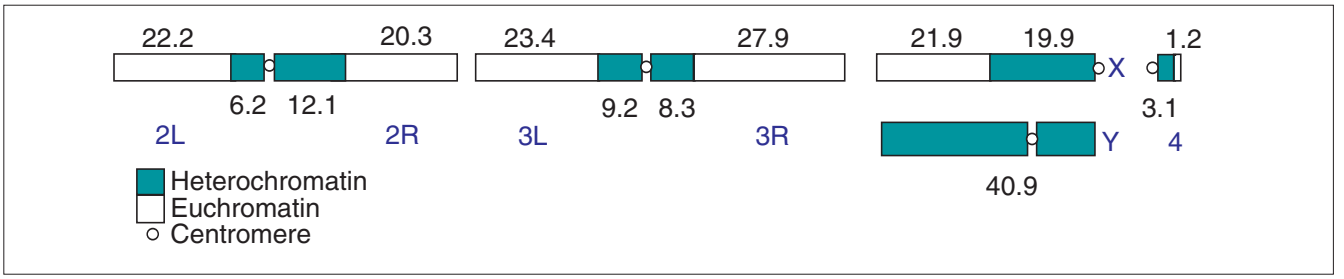

Figure 1
Chromosome structure of *Drosophila melanogaster*. The left and right arms of chromosomes 2 (2L, 2R) and 3 (3L, 3R), the small chromosome 4, and the sex chromosomes X and Y are shown (adapted from [12]). The numbers correspond to lengths in megabases. The euchromatic portions of the chromosome arms (white) correspond to the Release 3 euchromatic sequence described in Celniker *et al.* [47]. The lengths of the heterochromatic portions of the chromosome arms (green) are estimated from measurements of mitotic chromosomes [76]. The length of the heterochromatin on the X chromosome is polymorphic among strains and can comprise from one-third to one-half of the length of the mitotic chromosome. Our cytogenetic experiments show that Release 3 euchromatic sequence (white) extends into the centric heterochromatin by approximately 2.1 Mb (see Results).

(α-like G protein subunit), *Nipped-B* (morphogenesis), *rolled* (MAP kinase), poly(ADP-ribose) polymerase (chromatin structure), *bobbed* (ribosomal RNA), and the Y-linked fertility factors *kl-2*, *kl-3* and *kl-5* (dynein heavy chains) [33]. Genes have been localized on the cytogenetic map through analysis of chromosomal rearrangements and by fluorescence *in situ* hybridization (FISH) [25,40,41]. The genomic structures of several of these genes have been determined, and they differ from those of euchromatic genes. Their introns and regulatory regions are composed of clusters of partial and complete transposable elements, and some introns are hundreds of kilobases in length [42-46].

The Release 1 whole-genome shotgun (WGS) assembly of *Drosophila* [12] included 3.8 Mb of short, unmapped scaffolds representing heterochromatic sequence. The most recent version, WGS3 [47], assembled a total of 137.7 Mb of the *Drosophila* genome, using an improved assembly algorithm and the same trace data used for Release 1. A high-quality sequence of 116.9 Mb that spans the euchromatic portions of the chromosome arms is reported in [47]. For the sake of consistency, we refer to this 116.9-Mb sequence as the 'Release 3 euchromatic sequence' even though, on the basis of the cytological criteria for defining the boundary between euchromatin and heterochromatin described below, we believe this sequence extends into the centric heterochromatin of each chromosome arm. The annotation of genes and transposable elements in this approximately 2 Mb of heterochromatin-derived DNA are reported in [32,48]. Here, we characterize and annotate the 20.7 Mb of WGS3 sequence, the 'WGS heterochromatic sequence', that is not represented in the 116.9 Mb Release 3 euchromatic sequence.

Results

Annotation of WGS3 heterochromatic sequences

WGS3 was aligned to the 116.9 Mb Release 3 sequence spanning the euchromatin [47]. The WGS sequence that extends beyond the Release 3 euchromatic sequence (or otherwise fails to align) is an unfinished, draft-quality assembly of a total of 20.7 Mb of the heterochromatic portion of the *D. melanogaster* genome. The WGS3 heterochromatic sequence is distributed in 2,597 scaffolds. It includes portions of five scaffolds that overlap with and extend the Release 3 euchromatic sequences of chromosome arms X, 2L, 2R, 3L and 3R. The scaffolds range from 1 kb to 712 kb (Figure 2) and include 1,170 sequence gaps accounting for 3.7 Mb (18%). Some scaffolds, particularly those under 2 kb, may map within sequence gaps of larger scaffolds.

The WGS3 heterochromatic sequence was aligned to the corresponding portion of the Release 2 sequence. Although there are local differences between the two assemblies, the order and orientation of sequences are well conserved. The most significant difference between them is that WGS3 heterochromatic scaffolds are often considerably longer than the corresponding Release 2 scaffolds. For example, a 252-kb WGS3 scaffold extends over 13 smaller Release 2 scaffolds, and includes all seven exons of the heterochromatic gene *rolled* in the correct order and orientation (Figure 3). *rolled* maps to bands h40-h41 on the right arm of chromosome 2 [49].

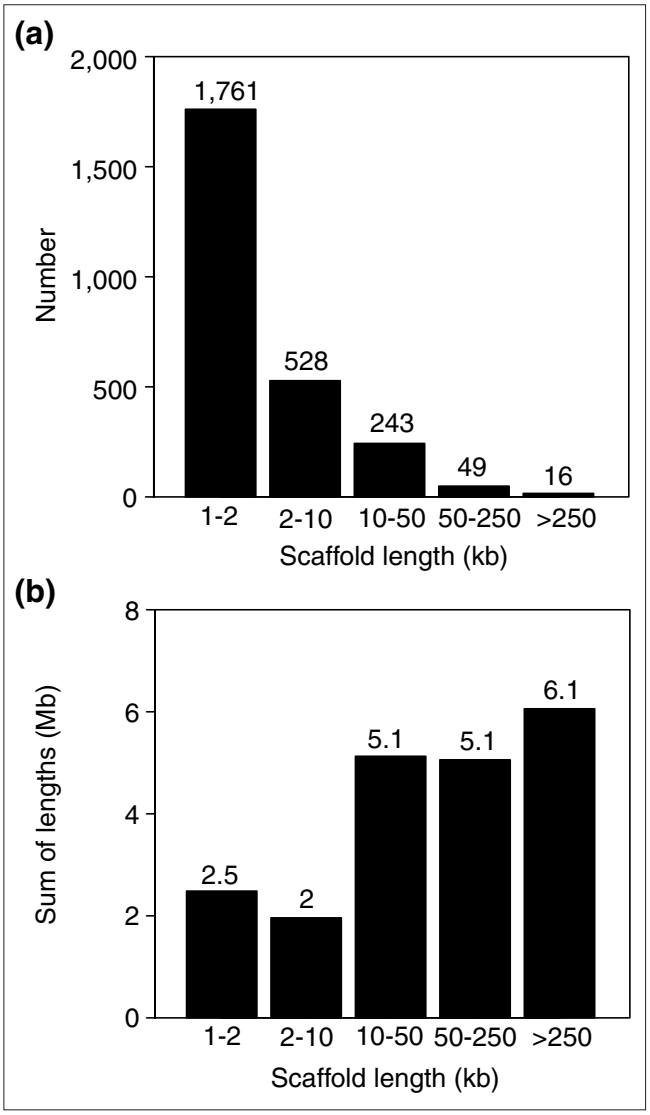

Figure 2
Distribution of scaffold lengths in the WGS3 heterochromatic sequence. **(a)** Histogram of the number (indicated above each bar) of sequence scaffolds in each of the indicated size ranges (kb). **(b)** Histogram of the sequence total (Mb; indicated above each bar) represented in the scaffolds in each of the indicated size ranges (kb).

We annotated the 12.0 Mb of WGS3 heterochromatic sequence in the 85 scaffolds longer than 40 kb, plus a 133-kb sequence at the centric end of the Release 3 euchromatic sequence of the X chromosome that was not annotated by Misra *et al.* [48]. We arbitrarily excluded the 8.7 Mb of

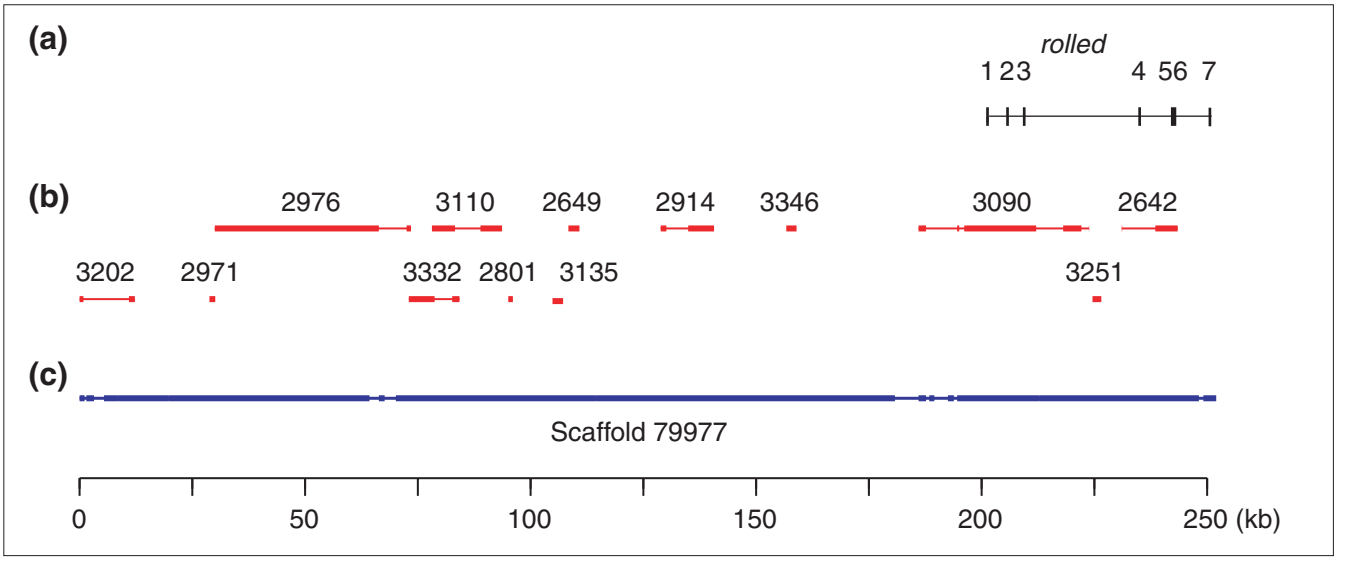

Figure 3
Comparison of WGS3 and Release 2 sequence assemblies of the *rolled* region. **(a)** Genomic organization of the *rolled* gene. Exons are shown as black boxes numbered 1 to 7, and introns are shown by the thin black line. All exons are present in a single WGS3 scaffold; exons 4 and 7 are absent from Release 2. **(b)** Thirteen Release 2 sequence scaffolds are shown as red bars. Thick portions of bars show regions aligned to WGS3, and thin portions show unaligned regions corresponding to sequence gaps. Scaffolds are labeled with the GenBank accession numbers, all of which begin 'AE00' and end in the indicated four digits: for example, AE003202. **(c)** The 252-kb WGS3 heterochromatic sequence scaffold 211000022279977 (Scaffold 79977), shown by the blue bar, links the 13 Release 2 scaffolds. The thin portions of the bar represent sequence gaps.

WGS3 heterochromatic sequence in the 2,512 scaffolds shorter than 40 kb from detailed annotation. Preliminary analysis suggested that gene identification in these small scaffolds is hampered in part by the separation of exons onto different scaffolds. This is illustrated by our annotation of seven 'super-scaffolds' that were constructed by linking together 25 short WGS3 heterochromatic scaffolds using cDNA evidence (see below). Thus, the scaffolds shorter than 40 kb do contain genes, but a reliable annotation of these sequences will require further analysis.

Identifying genes within heterochromatin presents challenges not encountered when annotating euchromatic sequences. Open reading frames (ORFs) in transposable elements can interfere with the identification of single-copy protein-coding genes, particularly when transposable elements are nested within introns. Also, heterochromatic genes can have large introns separating relatively small exons [42-46]. Therefore, transposable element and low-complexity sequences were masked; then the lengths of masked regions and sequence gaps were reduced to a maximum of 70 bp, which is the median intron length of euchromatic protein-coding genes in *Drosophila* [48,50].

We annotated the masked scaffolds using the computational annotation pipeline developed by Mungall *et al.* [51] and the annotation tool Apollo [52]. The pipeline generates, stores and filters alignments of expressed sequence tags (ESTs), cDNAs, and the results of protein similarity searches and gene-prediction algorithms. Apollo displays the filtered results of the pipeline in tiers of evidence, and allows human curators to evaluate and use the evidence to construct gene models. The guidelines used to define gene models in the euchromatin [48] were modified slightly, to deal with the unique properties of heterochromatic sequence (see Materials and methods). We generated 351 preliminary gene models on the masked scaffolds.

Next, the preliminary gene models were re-curated on the unmasked WGS3 heterochromatic sequence scaffolds. The unmasked scaffolds were run through the computational annotation pipeline, and the preliminary gene models were aligned to this genomic sequence. Twenty-five preliminary gene models could not be aligned to the unmasked scaffolds using sim4 [53]. After further examination, 11 of these were accepted as curated gene models, six were similar to transposable elements and were rejected, and eight could not be reconciled with the unmasked sequence. A higher-quality genomic sequence may be required to verify these eight models, and they have not been included here. The remaining preliminary gene models aligned in a consistent manner. After re-examination of the evidence, a total of 293 preliminary gene models were accepted, including 287 protein-coding gene models and five non-protein-coding gene models.

Because they are present on WGS3 heterochromatic scaffolds shorter than 40 kb, a number of previously known Y-linked protein-coding genes were missed by our analysis.

We annotated six of these Y-linked genes (*kl-2*, *kl-3*, *kl-5*, *Ory*, *Pp1-Y2* and *Ppr-Y*). The WGS sequence data were generated from clone libraries made from mixed-sex populations, so the male Y chromosome is represented by a four-fold lower density of sequence reads than the autosomes. In addition, Y-linked genes can have very large introns [46]. Consequently, sequences on the Y chromosome are represented by shorter scaffolds in the WGS assemblies [54]. In fact, five of the six Y-linked genes that we annotated were first characterized by analysis of short WGS scaffolds [55,56]. We used cDNA sequences to identify short WGS3 scaffolds bearing fragments of each of the six genes, concatenated these scaffolds into larger scaffolds ('super-scaffolds'; see Materials and methods), and annotated the resulting sequences to produce gene models. We produced and annotated one additional super-scaffold (linked_7) using EST evidence. The seven super-scaffolds contain 10 protein-coding gene models and one non-protein-coding gene model.

Evidence for the gene models

We generated 297 protein-coding gene models, and six non-protein-coding gene models. These include 30 previously known and molecularly characterized protein-coding genes (see Supplementary Table 1 in the additional data file). The protein-coding gene models are supported by four classes of evidence: alignment to *Drosophila* ESTs; alignment to *Drosophila* full-insert cDNA sequences; protein similarity; and prediction by gene-finding algorithms. There are 134 gene models (45%) that overlap at least one aligned EST, and 58 of these (20%) are further supported by an aligned cDNA sequence. There are 167 models (54%) with protein similarity evidence and 241 models (81%) supported by gene prediction. The numbers of gene models supported by all combinations of the classes of evidence is diagrammed in Figure 4. The 28 models (9%) supported only by EST evidence require further validation, because cDNA libraries can be contaminated with various artifacts, including sequences from unprocessed transcripts. The 89 models (30%) supported only by gene prediction are included in our curated set because they encode peptides of at least 100 amino acids with no significant similarity to proteins encoded by transposable elements (see below). Finally, the 15 models (5%) supported only by protein similarity have BLAST or TBLASTX expect scores (E-values) of less than 1×10^{-10} to sequences in *Drosophila* or other species.

We excluded gene models with greater than 95% identity over 50 bp to known transposable elements. We then examined nine remaining preliminary gene models with a TBLASTX E-value of less than 1×10^{-9} when compared to a database of transposable elements [32], and we retained all but one of them in our annotation. Although these gene models have weak similarity to proteins encoded by transposable elements, the extent of similarity is insufficient to suggest that they represent *bona fide* transposable element ORFs. Two of these models represent the genes *kl-5* and *scro*, and the region of similarity in *scro* corresponds to the homeodomain of this transcription factor gene. To further verify that the WGS3 heterochromatic annotations do not show significant homology to transposable elements, we compared the distributions of WGS3 heterochromatic and Release 3 euchromatic gene models with TBLASTX similarity to transposable elements over a range of expect scores (Figure 5). This analysis shows that the percentage of gene models with weak similarity to transposable elements is comparable in the WGS3 heterochromatic annotation and the Release 3 euchromatic annotation.

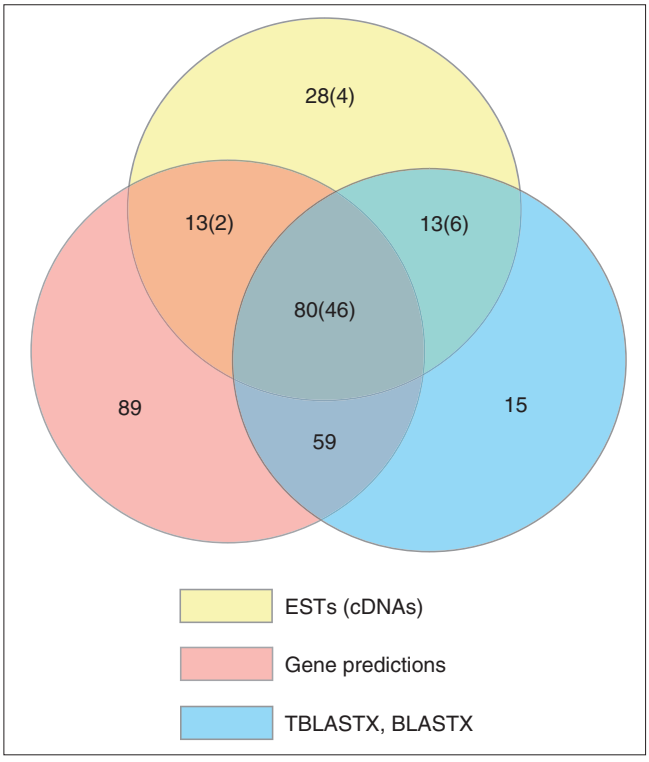

Figure 4
Evidence supporting the gene models. The diagram shows the numbers of curated gene models supported by evidence in three classes. Sequence alignments to *Drosophila* ESTs and, in parentheses, full-insert cDNA sequences were determined using sim4 (yellow circle). Gene predictions were made using Genie and Genscan (red circle). Similarities to known and predicted genes and proteins in *Drosophila* and other organisms were determined using BLASTX and TBLASTX (blue circle). The intersections in the diagram show the number of gene models that are supported by multiple evidence types. For example, there are 80 models supported by all three types of evidence, and 46 of these are represented by full-insert cDNA sequence. The 89 gene models supported only by gene prediction include 22 models that were predicted only in the masked sequence.

The WGS3 heterochromatic gene models are supported by fewer data than the Release 3 euchromatic gene models. In the WGS3 heterochromatin, 45% of gene models have an overlapping EST, compared to 78% in the Release 3 euchromatin. Twenty percent of gene models in the WGS3 heterochromatin are based on full-insert sequences of

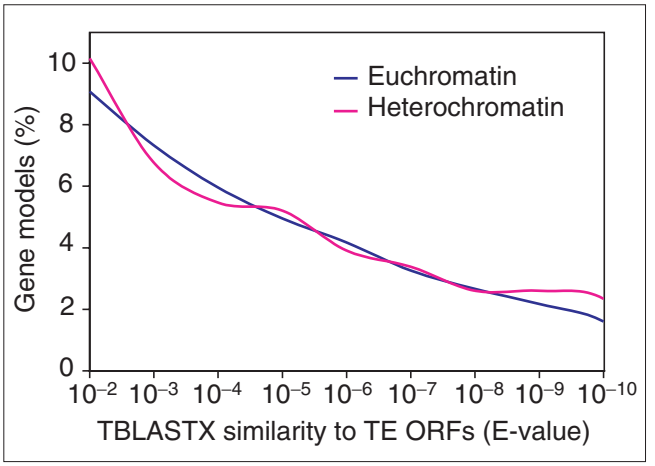

Figure 5
Gene models with weak similarity to transposable elements (TE). The percentage of gene models with TBLASTX similarity to known transposable elements at E-values from 1×10^{-2} to 1×10^{-10} is shown. The data are very similar for the Release 3 euchromatic (blue line) and WGS3 heterochromatic (pink line) annotations. The eight (2.7%) curated gene models in the WGS3 heterochromatin with E-values $\leq 1 \times 10^{-10}$ were examined further, as described in the text.

cDNAs, compared to over 70% in the Release 3 euchromatin. However, this observation is biased because half of the cDNAs in the *Drosophila* Gene Collection were selected for full-insert sequencing based on EST alignments to Release 2 gene models [57], and the WGS3 heterochromatin annotation preserves few of the Release 2 models and adds many new models. Many WGS3 heterochromatic gene models are based solely on gene predictions. Despite generating a large number of models, gene-finding algorithms were less successful at predicting heterochromatic genes than euchromatic genes. For example, only 72% of the heterochromatic gene models are supported by Genscan predictions, as opposed to 96% of the Release 3 euchromatic gene models.

Finally, we annotated six non-protein-coding RNA genes. WGS3 scaffold 211000022279294 contains rDNA sequences, including two complete copies each of the 5.8S and 2S rRNAs, a truncated 18S rRNA sequence that extends into a sequence gap, and a truncated 28S rRNA sequence that extends beyond the end of the scaffold. This region probably represents a portion of one of the two *bobbed* loci, which map to the X and Y chromosomes [58,59]. Two additional non-protein-coding gene models were annotated by similarity to euchromatic genes (see Supplementary Table 1 in the Additional data file).

Comparison to the Release 2 annotation

The annotation of WGS3 heterochromatic sequence has increased both the number and quality of gene models in the heterochromatin, relative to the corresponding portion of Release 2 (see Supplementary Table 2 in the additional data files). During the curation of the 297 protein-coding gene models in WGS3 heterochromatic sequence, 79 gene models from the corresponding portion of Release 2 were deleted, and 250 new gene models were created. Many of the deleted Release 2 annotations represent ORFs that overlap transposable elements. In annotating WGS3 heterochromatin, 10 Release 2 gene models were merged into five new models, one Release 2 model was split into two new models, and one Release 2 annotation was split into two models, one of which was merged with another model. Only 30 of the 130 Release 2 protein-coding gene models were preserved intact in the new annotation; 21 previous models were preserved with modifications. Thus, a much higher fraction of Release 2 gene models was modified in the WGS3 heterochromatin annotation than in the Release 3 euchromatic annotation, in which nearly two-thirds of predicted ORFs were unchanged [48].

As in the annotation of the Release 3 euchromatic sequence, the increased numbers of ESTs and cDNA sequences available for alignment to the WGS3 heterochromatic sequence resulted in significant improvements in the annotation of untranslated regions (UTRs), alternative transcripts, and intron-exon structures (see Supplementary Table 2 in additional data file). Twice as many genes in WGS3 heterochromatin have annotated 5′ and 3′ UTRs as in the corresponding portion of Release 2. The average 5′ UTR length is 258 bp and the average 3′ UTR length is 335 bp, both of which are close to the average UTR length of Release 3 euchromatic genes. There are 49 gene models with more than one transcript; only three were annotated in the corresponding portion of Release 2. There are 377 predicted protein-coding transcripts encoding 1,096 distinct exons, three times as many as were annotated in Release 2. The average number of introns per gene model increased from 2 to 2.7, but remains below the 3.6 average in the Release 3 euchromatin. The average length of introns increased significantly, from 892 bp in the corresponding portion of Release 2 to 3,743 bp in WGS3 heterochromatin. The longest annotated intron in WGS3 heterochromatin is 119,217 bp, dwarfing a 17,613 bp intron that was the longest in the corresponding portion of Release 2. Finally, only two introns longer than 10 kb were annotated in Release 2, but 76 gene models in WGS3 heterochromatin have introns longer than 10 kb. Whereas the majority of annotated introns in both the WGS3 heterochromatin and the Release 3 euchromatin are in the range of 50-70 bp, there are clearly more long introns in heterochromatic genes.

Annotations of selected regions

The second largest WGS3 heterochromatic scaffold annotated is 594 kb long (Figure 6) and incorporates five separate Release 2 scaffolds. The sequence of this scaffold is contained within a previously described BAC contig [60,61]. Because the scaffold contains two well studied autosomal heterochromatic genes, *light* and *concertina*, analysis of its sequence

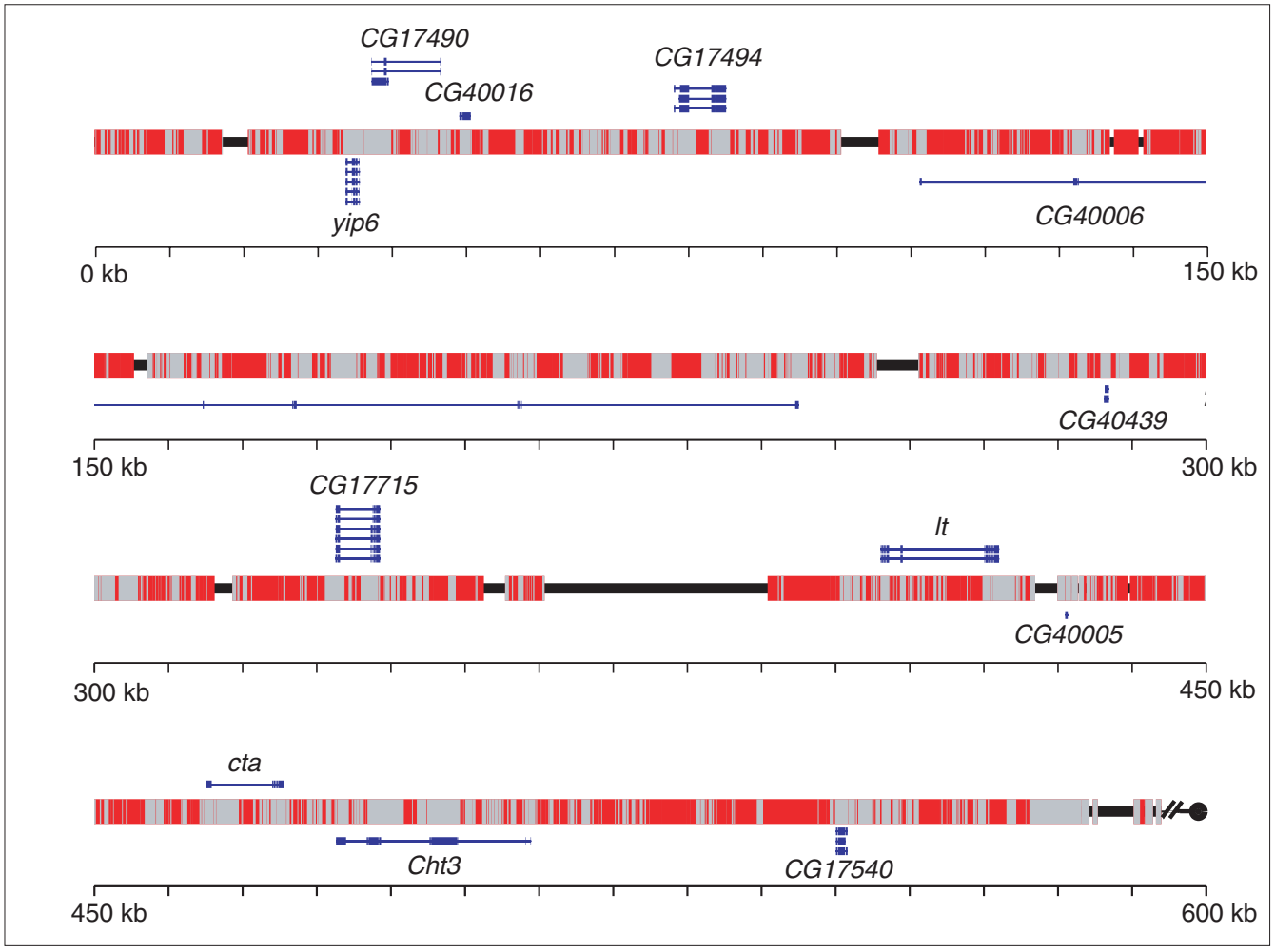

Figure 6
Annotation of the *light* (*lt*) region. The 594-kb WGS3 scaffold 211000022280798 and twelve curated gene models are shown. The WGS3 sequence is shown as a bar with sequence gaps (black), transposable elements and simple repeats that were masked and removed during the annotation process (red), and presumed single-copy sequences that remained after masking (gray) indicated. Gene models are shown as blue bars with exons (thick) and introns (thin) indicated. Those above the line are transcribed on the forward strand, and those below the line are transcribed on the reverse strand. The average density of curated genes is one per 50 kb, about six- to sevenfold lower than the density in the euchromatin [12,48]. Only the *lt* and *cta* genes are identified by genetic analysis. Seven gene models were described in the Release 1 annotation [12]. This annotation provides a more accurate view of the structures of nearly all of the gene models and determines their relative locations. cDNA sequence alignment allowed us to merge two Release 2 gene models, *Chitinase 1* and *3,* into a single gene *Cht3* with multiple chitin-binding and catalytic domains. Two of the three new curated genes (*CG40006, CG40016*) are represented by multiple ESTs and cDNAs. *CG40005* is based solely on BLAST evidence; its similarity to the adjacent *cta* gene suggests a possible sequence assembly artifact. On the basis of the masking results, known transposable element sequences account for 302 kb (51%) of the sequence scaffold.

provides an opportunity to assess the accuracy of both the WGS3 sequence assembly and the annotation process over an extended heterochromatic region. This scaffold maps within band h35 on 2Lh, and its orientation with respect to the centromere is known [49,61,62]. The scaffold lies approximately 100 kb proximal to the end of the Release 3 euchromatic sequence of 2L [47] as determined by genomic Southern blots using single-copy probes derived from the ends of the scaffold and the Release 3 2L sequence [61]. The structure of the scaffold is entirely consistent with previous 2Lh mapping studies in that the order of six single-copy sequences and the predicted *Not*I, *Pme*I, and *Sfi*I restriction maps of the scaffold (except for four regions corresponding to sequence gaps) are the same as those observed in the corresponding BAC contig and on genomic Southern blots. We annotated 12 protein-coding genes within the scaffold.

The annotation of the *rolled* gene illustrates how masking the WGS3 heterochromatic sequence improved Genscan performance and simplified curation (Figure 7). cDNA sequences define two alternative *rolled* transcripts that differ in their 5′ UTRs. All *rolled* exons defined by the cDNA

Figure 7
Annotation of the *rolled* gene. Results from the computational annotation pipeline for the portion of WGS3 scaffold 211000022279977 containing the *rolled* gene are displayed in Apollo. Evidence (black panels) used to annotate gene models (light blue panels) is shown. Evidence for gene models includes alignments of BLASTX results (red), cDNA sequences (green), and results of gene prediction (lavender). The curated structure of two *rolled* transcript models is defined by cDNA sequences. The Release 2 annotation (blue) did not include a complete *rolled* gene model. The predicted start (green) and stop (red) codons are indicated in the gene models. **(a)** In the annotation of the WGS3 masked sequence, in which transposable elements were removed, the Genscan prediction and the BLASTX results are consistent with the curated gene structure. The BLASTX evidence in the 3′ intron of *rolled* identifies an unmasked transposable element (yellow arrow). **(b)** In the unmasked WGS3 sequence, which includes known transposable elements (purple), Genscan fails to predict the first five exons of *rolled*, predicts two gene models within transposable elements, and adds three spurious exons to an inaccurate *rolled* gene model.

sequences align to the WGS3 heterochromatic scaffold and the masked scaffold in the correct order and orientation. In the annotation of the unmasked WGS3 scaffold, transposable elements within *rolled* introns interfered with Genscan so that a very poor *rolled* gene prediction was generated, and protein similarity results of transposable elements complicated the evidence. In the annotation of the masked scaffold, Genscan predicted a single gene model that includes all but one *rolled* coding exon, and protein similarity evidence supported the exon that Genscan missed.

Transposable elements

We carried out a preliminary analysis of the transposable element sequences found in the WGS3 heterochromatic sequence. We used a database of transposable elements [32] and RepeatMasker [63] to measure the amount of sequence that was derived from each transposable element family. Many of the sequences we identified represent only portions of elements; such fragmentary elements are often generated when transposable elements insert into one another to form complex nests [32]. Despite this complication, we were able to estimate the contribution of each transposable element class.

The most striking observation is the high fraction of the WGS3 heterochromatic sequence that is derived from transposable elements. We found that 52% of the 20.7-Mb WGS3 heterochromatic sequence had similarity to known transposable elements. Using similar analyses, transposable elements account for just 5.0% of the Release 3 euchromatic sequence; a slightly lower value of 3.9% was obtained in the analyses reported in [32], which required a higher level of sequence conservation. There were also some differences in the relative contributions made by different classes of elements in heterochromatin and euchromatin. LTR elements represent 61% of euchromatic transposable elements and approximately 78% of heterochromatic elements. LINE elements represent 24% of the euchromatic and 17% of the heterochromatic transposable element sequence. TIR elements represent 15% in euchromatin and 5% in heterochromatin. No FB elements were identified using RepeatMasker; a more targeted search identified 12 kb (0.1%) of FB element sequence.

Although we found a much higher density of transposable element sequences in the heterochromatin than in the euchromatin, it is likely that we missed many heterochromatic transposable elements. In fact, we found ORFs with similarity to transposable elements, such as those encoding transposases, outside those regions we annotated as transposable elements (see Figure 7a, for example), suggesting the existence of novel transposable element families. Finally, many of the sequence gaps within scaffolds probably correspond to regions of the genome with very high transposable element density. Thus, our analysis almost certainly represents an underestimate of the total transposable element content of the WGS3 heterochromatic scaffolds. As repetitive elements are difficult to assemble using the WGS strategy, an accurate estimate of their contribution to the heterochromatic sequence awaits a more finished version of the sequence.

Cytological boundaries of centric heterochromatin

Although there is no universally accepted definition of heterochromatin, the most reliable classification of the centric heterochromatin is cytological. Therefore, we consider the sequences located within the bands (h1-h61) of the cytological map defined on mitotic chromosomes to be heterochromatic. In order to correlate heterochromatin cytology with the genome sequence, we mapped the boundaries of the euchromatin and centric heterochromatin on chromosome arms X, 2L, 2R, 3L and 3R by FISH of BAC-derived probes to mitotic chromosomes (see Materials and methods). BACs were selected from the centric ends of the essentially finished Release 3 sequence contigs that span the euchromatin [47], and the positions of the resulting hybridization signals on the cytological map were determined (Figure 8, Table 1). These data show that the large Release 3 sequence contigs extend into the centric heterochromatin on these five chromosome arms. Although BACs were localized to chromosome 4 (data not shown), the cytology of this chromosome is too poor to permit a clear BAC-based mapping of the boundary of the centric heterochromatin. The Y chromosome has no boundary, because it is entirely heterochromatic.

For the purposes of defining the gene content of the heterochromatic portion of the genomic sequence, we provisionally designate the distal ends of the indicated BACs (Table 1) as defining the boundaries of the centric heterochromatin within the finished genomic sequence. However, the transition from euchromatin to heterochromatin appears to be gradual rather than sharp, and the resolution of these cytological mapping experiments appears to be on the order of 100 kb. Therefore, we have approximately localized the heterochromatin-euchromatin boundaries with respect to the genomic sequence, and have defined precise boundaries here simply as a convenience in discussing the genome annotation data. By this definition, the Release 3 sequence that spans the euchromatin includes 2.1 Mb of sequence in centric heterochromatin, and this sequence includes 150 curated genes described in Misra *et al*. [48]. These genes are in addition to those identified in our analysis, which was restricted to the WGS3 heterochromatic scaffolds described above.

The BACs that we localized on the cytological map of the mitotic chromosomes have also been positioned using *in situ* hybridization to salivary gland polytene chromosomes [47,60]. Although banding in the proximal regions of the polytene chromosomes is not as distinct as in the euchromatic regions, comparison of the two datasets shows the approximate extent of overlap between the mitotic and polytene cytogenetic maps (Table 1). The boundary of the centric heterochromatin of the X chromosome at the distal edge of band h26 corresponds approximately to polytene division 20C, band h35 on chromosome arm 2L corresponds to polytene division 40A, band h45 on 2R corresponds to polytene division 41F, band h47 on 3L corresponds to polytene division 80A, and band h58 on 3R corresponds to polytene division 82C.

Discussion

Our work has resulted in a substantially improved view of the sequence, organization, and gene content of the

Drosophila heterochromatin. The 20.7-Mb heterochromatic WGS sequence we describe here, together with the essentially finished 116.9-Mb euchromatic sequence described in Celniker *et al.* [47] and Misra *et al.* [48], constitute the 137.7-Mb Release 3 version of the annotated *D. melanogaster* genomic sequence.

We have demonstrated the efficiency and utility of WGS sequencing in assembling the single-copy and middle-repetitive regions within the heterochromatic portion of a complex genome. WGS sequencing samples at random the entire portion of the genome that is clonable in 2-kb segments. The ability to clone genomic regions not clonable in BACs or other large-insert vectors makes WGS sequencing essential to study the heterochromatic regions of complex genomes. We also describe a successful annotation strategy for these highly repetitive regions of the *Drosophila* genome.

The heterochromatic portion of the genome has a far higher content of repetitive sequences and a lower gene density than the euchromatin. Nevertheless, the number and importance of heterochromatic genes are significant. Although the gene models are supported by fewer data than those in euchromatin, our analysis has identified 297 predicted protein-coding genes and six non-protein-coding genes in the WGS3 heterochromatic sequence, and suggests that

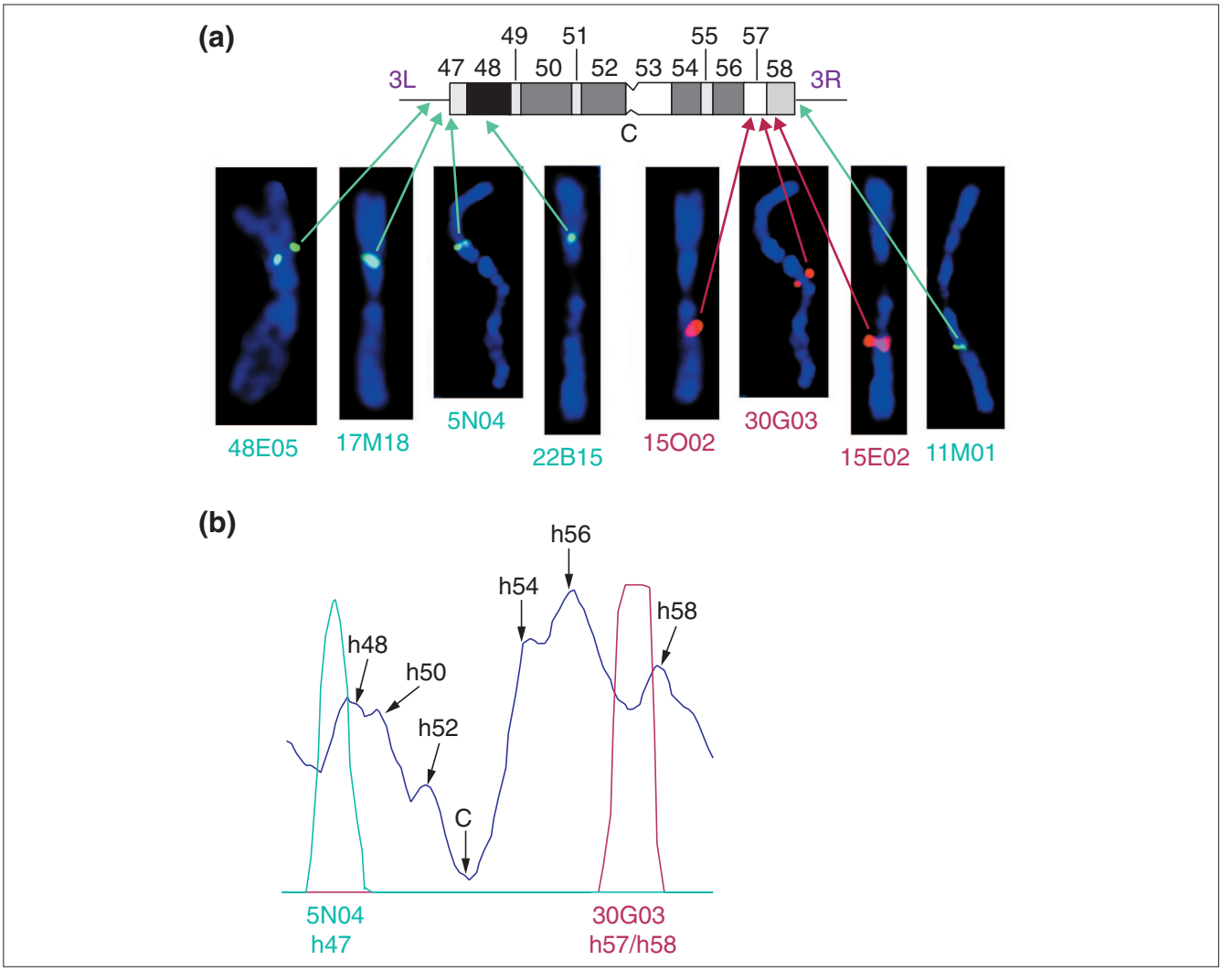

Figure 8
The boundaries of the centric heterochromatin defined by FISH. BACs near the centric ends of the Release 3 chromosome arm sequence spanning the euchromatin [47] were localized by FISH to mitotic chromosomes to correlate the cytological boundaries of the centric heterochromatin with the genomic sequence (see Materials and methods). **(a)** Results for chromosome 3 are shown. Locations of BACs on the cytogenetic map (bands h1-61) are indicated by arrows. The left (3L) and right (3R) arms, and the centromere (C), are indicated. BAC names are indicated below each image, and images are oriented with the left arm at the top. See Table 1 for complete BAC names and additional information. **(b)** An example of the quantitative analysis used to determine BAC locations (red and green) relative to the DAPI (blue) banding pattern is shown (see Materials and methods).

Table 1

Localization of BACs to the mitotic and polytene cytogenetic maps

BAC*	Mitotic		Polytene location§
	Location†	Comments‡	
XL ↓			
BACR20N11	Distal to h26	Euchromatic	19F
BACR09F10	Distal to h26	Euchromatic	19F-20A
BACR23I18	Distal to h26	Euchromatic	19F-20A
BACR05K22	Just distal to h26	Euchromatic	20C
BACR02D03	h26, distal edge	Transition	20C
BACR08A09¶	h26, distal edge	Transition	20C
BACR05O07	h26	Heterochromatin	20D
CEN X			
2L ↓			
BACR10I13¶	h35, distal edge	Transition	40A
BACR13A13	h35	Heterochromatin	40A-B
BACR13E23	h35	Heterochromatin	40B
BACR27P22	h35	Heterochromatin	40A-B
BACR38I21	h35	Heterochromatin	40A
BACR01K02	h35#	Heterochromatin	
BACR20K08	h35/h36#	Heterochromatin	
CEN 2			
BACR11B22	h43-45	Heterochromatin	41D
BACR07J16	h45	Heterochromatin	41D-E
BACR11B14	h45	Heterochromatin	41D-E
BACR06P07	h45	Heterochromatin	41E
BACR02D22¶	h45	Heterochromatin	CC
2R ↑			
3L ↓			
BACR17M18¶	h47, distal edge	Transition	80A-B
BACR05N04	h47	Heterochromatin	N/A
BACR22B15	h48/49	Heterochromatin	80B
CEN 3			
BACR15O02	h57	Heterochromatin	82A-B
BACR30G03	h57-58	Heterochromatin	82A-B
BACR15E02¶	h57	Heterochromatin	82A-C
BACR11M01	Distal to h58	Euchromatin	82E
BACR20D10	Distal to h58	Euchromatin	82E-F
3R ↑			

*All BACs are from the RPCI-98 library [60,72]. Arrows indicate orientation from euchromatin into heterochromatin; the order of BACs in each group reflects their relative positions within the finished genomic sequence, but is not intended to imply overlap. †Locations relative to the heterochromatic bands (h1-61) [5] are indicated. ‡Positions within mitotic euchromatin, heterochromatin, and the transition between them are indicated. §Localizations on the polytene chromosome map are indicated [47,60]. ¶BACs that define the approximate boundaries of the centric heterochromatin. #Locations of two BACs reported in Yasuhara *et al.* [61]. CC, chromocenter.

approximately 150 genes in the Release 3 euchromatic sequence annotated by Misra *et al.* [48] are also located in the cytologically defined heterochromatin. The organization and composition of heterochromatic and euchromatic genes appear to differ; heterochromatic genes in general contain larger transcription units with some unusually large introns, and introns consist predominantly of transposable elements. Although heterochromatic genes appear to differ from euchromatic genes in some aspects of gene structure, they do not appear to be segregated in any obvious way based on function. The predicted products of the approximately 450 predicted heterochromatic genes represent diverse biochemical activities that are likely to be involved in a wide range of essential functions.

Annotation of the 2.9-Mb *Adh* region identified 55 vital loci and 218 protein-coding gene models (25% essential genes) in a presumed typical euchromatic region [64]. Here, we describe 447 protein-coding gene models in the heterochromatin, including 150 models annotated in [48], but there are only 32 identified heterochromatic genes required for viability or fertility (7.2%) (see Background). This difference may or may not be significant, given that different euchromatic regions appear to have different ratios of essential genes to total genes [33]. There are several possible reasons for the apparent discrepancy between our results and the genetic analyses. First, saturating genetic screens have not been reported for all of the heterochromatin, so the number of essential loci is underestimated. Second, the centric heterochromatin is defined more narrowly in the genetic analyses than in our analysis. For example, the WGS3 heterochromatic sequence includes *suppressor of forked* (*su(f)*) and the dicistronic *stoned* locus (*stnA+stnB*) (see Supplementary Table 1 in the additional data file), which map near the boundary on the X chromosome and have not been described previously as heterochromatic loci. Third, we may have predicted too many genes. In our annotation, predicted proteins encoded by gene models without full-length cDNA sequence data are shorter on average (297 amino acids) than those encoded by gene models based on full-length cDNA sequences (376 amino acids). Thus, we expect additional cDNA sequence data will result in merges of adjacent gene models, reducing the number of predicted genes. In addition, gene models with low levels of supporting evidence may not represent valid genes. In this context, it is important to note that annotation of the WGS3 scaffolds shorter than 40 kb will probably result in the identification of more heterochromatic genes. Thus, resolution of this issue will require further experimentation.

We have described assembled sequences representing 22.8 Mb of the heterochromatin, including 20.7 Mb in 'WGS3 heterochromatic sequence' and approximately 2.1 Mb in 'Release 3 euchromatic sequence'. Because most of the WGS3 scaffolds have not been mapped to chromosomes, we

do not yet know how the assembled sequences are distributed within the 59 Mb of heterochromatin in the female genome and the additional 41 Mb of heterochromatin in the male genome. In addition to the 20.7-Mb sequence assembled in scaffolds, WGS3 includes 181,686 sequence traces clustered into 35,039 'degenerate scaffolds' representing repetitive sequences that were not assigned to unique locations in the assembly (E.W. Myers *et al.*, unpublished work). These sequences include transposable elements and satellite sequences (unpublished data). Satellite sequences represent approximately 20% of the genome and can be cloned in plasmids, but such clones are inefficiently recovered and unstable [22,65]. Thus, we do not know what fraction of the remaining heterochromatic sequence is sampled by these additional, unassembled sequence traces. Therefore, WGS data cannot be used to estimate accurately the fraction of the heterochromatin that can be recovered in stable plasmid clones.

Improvements to the annotation

The annotation of the heterochromatic portion of the *Drosophila* genome described here is a work in progress. We have annotated protein-coding genes, and summarized preliminary observations on non-protein-coding genes and transposable elements. Our analysis was limited by the high repeat content of heterochromatin and by the unfinished quality of the WGS3 heterochromatic sequence. Our decision to delay annotation of the scaffolds shorter than 40 kb has probably resulted in failure to identify some genes, especially on the Y chromosome. Despite these limitations, the protein-coding gene annotations are generally reliable, as demonstrated by the identification of previously known heterochromatic genes, and the alignment of cDNA sequences to the draft genomic sequence and the annotated gene models. Nevertheless, the quality of the annotations will be greatly improved by the addition of more full-length cDNA sequences of heterochromatic genes and by comparative analysis using the mosquito [66] and *D. pseudoobscura* [67] WGS sequences.

Future analyses of the differences between euchromatic and heterochromatic sequence may lead to improvements in the performance of computational gene-prediction algorithms on heterochromatic sequence. Our observations that the gene-prediction tools Genie [68] and Genscan [69] performed relatively poorly in identifying heterochromatic genes suggests that these programs could be modified to improve their performance on heterochromatic sequence. Processing the genomic sequence, by masking repeats and reducing the distances between potential coding exons, improved the performance of the gene-prediction tools, and improved gene identification during subsequent re-annotation of the unmasked sequence. Optimization of these preprocessing steps should lead to improved performance. The annotated gene models that are supported by cDNA and/or high-quality TBLASTX matches provide a useful dataset for training and testing gene-prediction algorithms on heterochromatic sequence.

A collection of approximately 600 *P* transposable element insertions in heterochromatin has recently been generated [70] (A.Y. Konev , C.M. Yan, E. O'Hagan , S. Tickoo, G.H.K., unpublished data). These *P* element insertions will provide tools for the analysis of heterochromatic genes and manipulation of the heterochromatic portion of the genome. For example, *P*-element-mediated deletions of centric heterochromatin have been used to map genes and regions responsible for controlling gene expression and replication [40].

Improvements to the genomic sequence

We plan to improve the WGS3 heterochromatic sequence by filling sequence gaps and correcting assembly errors. The quality of the WGS assembly suggests a strategy for bringing these sequences to high quality: first, select a tiling path of 10-kb genomic clones from the WGS that span each scaffold; second, sequence each clone to high quality; third, assemble these 10-kb sequences to reconstruct the genomic sequence; and fourth, verify the assembly by comparison to cDNA sequence alignments and to restriction digests of genomic DNA, assayed if necessary on Southern blots. cDNA alignments will also be useful in linking separate scaffolds in cases in which the exons of a single gene lie in more than one scaffold. We gained extensive experience in each of these steps during our finishing of the euchromatic portion of the genome [47], and no new technology is required to bring the WGS scaffolds we have described here to finished quality.

Some regions of the heterochromatin are clonable in BACs. Three small BAC contigs from the genome physical map are located in the centric heterochromatin of chromosome arms 2L, 2R and 3L [60]. Draft sequences of BACs spanning these contigs were produced during the Release 1 phase of the genome-sequencing project [12,71]. The small BAC contig on 2L corresponds to the WGS3 scaffold that includes *light* and *concertina* (see Figure 6) [61], and the contigs on 2R and 3L also align to WGS3 scaffolds. We have also identified BACs containing the *rolled*, *PARP* and *SNAP25* genes in pilot STS content mapping experiments in the heterochromatin. Doubtless other regions of the heterochromatin will be represented in large-insert libraries, and BACs will be useful for linking and orienting short WGS3 heterochromatic scaffolds. However, we were unable to identify BACs containing the Y-linked genes *ccy* and *kl-5* in available *Drosophila* BAC libraries [60,72], perhaps due to high satellite DNA content. This suggests that not all heterochromatic regions assembled in WGS3 will be represented in BACs. We also do not yet know whether the highly repetitive nature of heterochromatin decreases the stability of sequences cloned in BACs. For these reasons, we favor a sequence-finishing strategy based on the 10-kb clones generated in the WGS.

Materials and methods

Genomic sequence alignments

The alignment of WGS3 to the essentially finished Release 3 sequence spanning the euchromatin is described in Celniker *et al*. [47]. WGS3 scaffolds that did not show significant alignment were defined to be heterochromatic. In addition, five WGS3 scaffolds aligned to the centric ends of the Release 3 euchromatic sequence and extended beyond them. We used sim4 [53] and MUMmer [73] to realign these five scaffolds, and the sequence extending beyond the aligned portions of the Release 3 euchromatic sequence contigs was extracted with a 60-bp overlap and included in the WGS3 heterochromatic sequence. Finally, a 133-kb region including BACR48D21 at the centric end of the Release 3 X-chromosome sequence [47] was not included in the Release 3 annotation of the euchromatin [48]. This sequence was included in the analysis of WGS3 heterochromatic sequence.

The WGS3 heterochromatic sequence scaffolds were aligned to the corresponding Release 2 sequence scaffolds using BLAST2 [74]. The alignment results were carefully examined, and the two assemblies were found to have few discrepancies.

Masking sequence scaffolds

We masked WGS3 heterochromatic sequences before annotation. We used RepeatMasker [63], with the default settings, to mask transposable elements [32] and low-complexity sequences. Next, we shortened all sequence gaps and repeat-masked regions to 70 bp.

Databases and tools

To annotate the WGS3 heterochromatic sequence, we used the computational analysis pipeline and databases described in Mungall *et al*. [51]. This pipeline aligns *Drosophila* ESTs, cDNA sequences, and other sequences in GenBank using sim4 [53], performs DNA and protein sequence similarity searches of the GenBank and SwissProt/TrEMBL databases using BLASTX and TBLASTX, and executes the gene-prediction algorithms Genie [68] and Genscan [69]. The results generated by the pipeline were filtered using the Bioinformatics Output Parser (BOP) [51], and the filtered results were curated using the tool Apollo [52].

Curation guidelines

To determine the intron-exon structure of gene models, curators visually inspected the alignment of computational evidence types to the WGS3 heterochromatic sequences using Apollo. Alternative transcripts supported by EST evidence and UTRs were annotated. The criteria used to curate the gene models in the WGS3 sequence were identical to those used in the annotation of the euchromatin [48], with two exceptions. First, computational results derived from Genscan were not judged by their scores, as it was empirically determined that low-scoring Genscan results often correctly predicted the intron-exon structures suggested by other evidence types such as cDNAs. Second, the predicted proteins of the WGS3 annotation were compared to a curated transposable element dataset [32]. Gene models with an alignment of at least 50 nucleotides with at least 95% identity to transposable elements were rejected.

Gene models produced on the masked scaffolds were mapped to the unmasked WGS3 sequence using sim4. Gene models that aligned with less than 95% identity to the unmasked sequence were not preserved. Gene models were checked to ensure that they did not overlap transposable elements and that they did not have major alterations of their intron-exon structure due to the presence of unmasked data. Gene models were refined using evidence that had not aligned to the masked WGS3 sequence.

Linking scaffolds with cDNAs

In WGS3, exons of the *kl-5* gene are distributed over four scaffolds. These scaffolds were concatenated and annotated to produce a *kl-5* gene model. Similar 'super-scaffolds' (linked_1 to linked_6) were constructed for the genes *kl-2*, *kl-3*, *kl-5*, *ORY*, *Pp1-Y2*, and *Ppr-Y*. An additional super-scaffold (linked_7) was constructed based on EST evidence. The super-scaffolds were constructed as follows, with each WGS3 scaffold indicated by the last five digits of the scaffold ID, and relative orientation indicated by F (forward) or R (reverse): linked_1 (80774R-78545F-78270R-78383F-79796R-78126F-78519R); linked_2 (80705F-80550F-80543F-80769R-80048R); linked_3 (80569R-79234F-79561F-80324F); linked_4 (80310F-80189F-80349R-80306F); linked_5 (78068R-78764R-80118R); linked_6 (80329F-78590F); and linked_7 (78279R-78567F). Because the gaps between individual scaffolds in a super-scaffold are not spanned by identified genomic clones, their lengths are undefined. These gaps are represented with a string of 1,000 Ns, following the convention established for the Release 1 genomic sequence [12].

Fluorescence *in situ* hybridization (FISH)

Mitotic chromosomes from third instar larval neuroblasts were obtained by standard procedures [25]. Slides were aged at room temperature for 24 h or for 2 h at 60°C, pretreated in 100 μg/ml RNaseA/2x SSC, pH 7 at 37°C for 30-60 min, immersed in a 70/95/100% ethanol series for 2 min each, then air-dried and kept on a slide warmer at 45°C. BAC DNA (1 μg) was labeled with biotin-16-dUTP (Roche) or digoxigenin-11-dUTP (Roche) by nick translation. Labeled BAC DNA (200-300 ng per 22 x 22 mm hybridization area) was precipitated at -80°C for 30 min or overnight at -20°C with salmon sperm DNA (2 μg), 1/10th vol sodium acetate, and 2 vol cold 100% ethanol. Probes were centrifuged for 20 min at 14,000 rpm and 4°C, washed in 70% ethanol, and briefly dried in a Sorvall SpeedVac. Hybridization mix (10-15 μl) was added to each dried pellet. All probes were initially hybridized in a solution containing 55% formamide/2x SSC, 20% dextran sulfate, 1% Tween-20, incubated overnight at 37°C, and washed in 55% formamide/2x SSC at 42°C for

20 min, followed by four washes in 2x SSC at 37°C (2 min each) and 1-3 washes in 0.1x SSC at 60°C (1 min each). For those BACs that demonstrated significant cross-hybridizaton to other chromosomal regions, FISH was repeated using higher stringency (60% formamide) hybridization solution and post-hybridization washes. In single-color hybridization experiments, biotinylated or digoxigenin-labeled BAC probes were detected using FITC-avidin (Vector Laboratories) or FITC anti-digoxigenin (Roche), respectively. Detections were performed overnight at 4°C or 1-3 h at room temperature. For multi-BAC (two-color) FISH, biotinylated probes were detected with FITC avidin and digoxigenin-labeled probes were detected with Rhodamine anti-digoxigenin (Roche). After incubation with avidin or anti-digoxigenin, slides were washed in coplin jars on a rotating shaker for three 5-min washes in 4x SSC/0.1% Tween-20. DNA was counterstained with Vectashield (Vector Laboratories) containing 1-5 μg/ml 4,6-diamidino-2-phenylindole (DAPI). The location of the BAC signals relative to the DAPI banding pattern on the heterochromatic map was determined by visual analysis in Photoshop (Adobe). In addition, an independent quantitative analysis using IP Labs (Signal Analytics, Vienna, VA) and a fluorescence quantitation script [70] was performed on each image. The fluorescence levels along lines drawn through the chromosome axis were plotted for the DAPI and BAC signals (Figure 8b), which produces a more precise localization than is possible by visual inspection of the images. A minimum of 10 prometaphase chromosomes were analyzed for each BAC, and localizations were determined by consensus.

Heterochromatin in the 'Release 3 euchromatic sequence'

The BACs at the boundaries between euchromatin and centric heterochromatin (see Table 1) identify 2.1 Mb (150 curated gene models) of heterochromatic sequence at the centric ends of the high-quality Release 3 sequence spanning the euchromatin [47]. The distal ends of the boundary BACs were identified using BAC end sequences. The sequence proximal to these positions is provisionally defined to be heterochromatic: X 21,561,835 to 21,912,668 bp (0.351 Mb, six genes); 2L 21,834,050 to 22,217,931 bp (0.384 Mb, 25 genes); 2R, 1 to 467,915 (0.468 Mb, 44 genes); 3L, 22,879,753 to 23,352,213 (0.472 Mb, 11 genes); and 3R, 1 to 378,655 (0.379 Mb, 64 genes).

Data deposition

WGS3 heterochromatic sequence and annotations will be deposited in GenBank and the FlyBase GadFly database [75], and the corresponding Release 2 sequences will be subsumed into the new sequence accessions. The WGS3 heterochromatic sequence, the annotations, and the evidence that supports them will be made available at FlyBase [71].

Additional data files

An additional data files containing Supplementary Tables 1 and 2 are available with the online version of this paper.

Acknowledgements

We thank Sima Misra and Casey Bergman for helpful discussions; Andrew Skora, Christopher Yan, and David Acevedo for assistance with FISH experiments; Robert Svirskas, John Tupy, Pavel Hradecky, Colin Wiel, Bruno Ribeiro, Marcelo Alvim and Maria Vibranovski for assistance with informatics; Erwin Frise, Eric Smith and Dave Hurley for computer systems support; and Catherine Nelson for editing the manuscript. This work was supported by Celera Genomics, the Howard Hughes Medical Institute, NSF grant MCB0213163 to B.T.W., fellowships from the CNPq and the Pew Latin American Fellows Program to A.B.C., NIH grant R01 HG00747 to G.H.K. and NIH grant P50 HG00750 to G.M.R. The work supported by P50-HG00750 was carried out under Department of Energy Contract DE-AC0376SF00098, University of California.

References

1. Heitz E: **Das Heterochromatin der Moose.** *I Jahrb Wiss Botanik* 1928, **69:**762-818.
2. John B: **The biology of heterochromatin.** In *Heterochromatin: Molecular and Structural Aspects*, Edited by Verma RS. Cambridge: Cambridge University Press; 1988:1-147.
3. Elgin SC, Workman JL: **Chromosome and expression mechanisms: a year dominated by histone modifications, transitory and remembered.** *Curr Opin Genet Dev* 2002, **12:**127-129.
4. Weiler K, Wakimoto B: **Heterochromatin and gene expression in *Drosophila*.** *Annu Rev Genet* 1995, **29:**577-605.
5. Gatti M, Pimpinelli S: **Functional elements in *Drosophila melanogaster* heterochromatin.** *Annu Rev Genet* 1992, **26:**239-275.
6. Sullivan B, Karpen G: **Centromere identity in *Drosophila* is not determined *in vivo* by replication timing.** *J Cell Biol* 2001, **154:**683-690.
7. McKee BD, Karpen GH: ***Drosophila* ribosomal RNA genes function as an X-Y pairing site during male meiosis.** *Cell* 1990, **61:**61-72.
8. Dernburg AF, Sedat JW, Hawley RS: **Direct evidence of a role for heterochromatin in meiotic chromosome segregation.** *Cell* 1996, **86:**135-146.
9. Karpen GH, Le MH, Le H: **Centric heterochromatin and the efficiency of achiasmate disjunction in *Drosophila* female meiosis.** *Science* 1996, **273:**118-122.
10. Moore DP, Orr-Weaver TL: **Chromosome segregation during meiosis: building an unambivalent bivalent.** *Curr Top Dev Biol* 1998, **37:**263-299.
11. Bernard P, Maure JF, Partridge JF, Genier S, Javerzat JP, Allshire RC: **Requirement of heterochromatin for cohesion at centromeres.** *Science* 2001, **294:**2539-2542.
12. 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.
13. Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, *et al.*: **Genetic definition and sequence analysis of *Arabidopsis* centromeres.** *Science* 1999, **286:**2468-2474.
14. Horvath JE, Schwartz S, Eichler EE: **The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome.** *Genome Res* 2000, **10:**839-852.
15. Horvath JE, Viggiano L, Loftus BJ, Adams MD, Archidiacono N, Rocchi M, Eichler EE: **Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11.** *Hum Mol Genet* 2000, **9:**113-123.
16. Haupt W, Fischer TC, Winderl S, Fransz P, Torres-Ruiz RA: **The centromere1 (CEN1) region of *Arabidopsis thaliana*: architecture and functional impact of chromatin.** *Plant J* 2001, **27:**285-296.
17. Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H: **The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 5.** *DNA Res* 2000, **7:**315-321.
18. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
19. Kotani H, Hosouchi T, Tsuruoka H: **Structural analysis and complete physical map of *Arabidopsis thaliana* chromosome 5**

including centromeric and telomeric regions. *DNA Res* 1999, **6:**381-386.
20. Carvalho AB: **Origin and evolution of the *Drosophila* Y chromosome.** *Curr Opin Genet Dev* 2002, **12:**664-668.
21. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF: **Genomic and genetic definition of a functional human centromere.** *Science* 2001, **294:**109-115.
22. Sun X, Le H, Wahlstrom J, Karpen GH: **Sequence analysis of a functional *Drosophila* centromere.** *Genome Res*, in press.
23. Heitz E: **Uber α- und β-Heterochromatin sowie Konstanz und Bau der Chromomeren bei *Drosophila*.** *Biol Zentbl* 1934, **54:**588-609.
24. Gall JG, Cohen EH, Polan ML: **Reptitive DNA sequences in *Drosophila*.** *Chromosoma* 1971, **33:**319-344.
25. Gatti M, Bonaccorsi S, Pimpinelli S: **Looking at *Drosophila* mitotic chromosomes.** *Methods Cell Biol* 1994, **44:**371-391.
26. Lohe AR, Hilliker AJ, Roberts PA: **Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*.** *Genetics* 1993, **134:**1149-1174.
27. Pimpinelli S, Berloco M, Fanti L, Dimitri P, Bonaccorsi S, Marchetti E, Caizzi R, Caggese C, Gatti M: **Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin.** *Proc Natl Acad Sci USA* 1995, **92:**3804-3808.
28. Le MH, Duricka D, Karpen GH: **Islands of complex DNA are widespread in *Drosophila* centric heterochromatin.** *Genetics* 1995, **141:**283-303.
29. Sun X, Wahlstrom J, Karpen G: **Molecular structure of a functional *Drosophila* centromere.** *Cell* 1997, **91:**1007-1019.
30. Losada A, Abad JP, Villasante A: **Organization of DNA sequences near the centromere of the *Drosophila melanogaster* Y chromosome.** *Chromosoma* 1997, **106:**503-512.
31. Miklos GL, Yamamoto MT, Davies J, Pirrotta V: **Microcloning reveals a high frequency of repetitive sequences characteristic of chromosome 4 and the beta-heterochromatin of *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 1988, **85:**2051-2055.
32. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, *et al.*: **The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective.** *Genome Biol* 2002, **3:**research0084.1-0084.20.
33. **FlyBase** [http://flybase.bio.indiana.edu]
34. Hilliker AJ: **Genetic analysis of the centromeric heterochromatin of chromosome 2 of *Drosophila melanogaster*: deficiency mapping of EMS-induced lethal complementation groups.** *Genetics* 1976, **83:**765-782.
35. Schupbach T, Wieschaus E: **Female sterile mutations on the second chromosome of *Drosophila melanogaster*. I. Maternal effect mutations.** *Genetics* 1989, **121:**101-117.
36. Marchant GE, Holm DG: **Genetic analysis of the heterochromatin of chromosome 3 in *Drosophila melanogaster*. II. Vital loci identified through EMS mutagenesis.** *Genetics* 1988, **120:**519-532.
37. Hilliker AJ, Appels R: **Pleiotropic effects associated with the deletion of heterochromatin surrounding rDNA on the X chromosome of *Drosophila*.** *Chromosoma* 1982, **86:**469-490.
38. Sinclair DA, Schulze S, Silva E, Fitzpatrick KA, Honda BM: **Essential genes in autosomal heterochromatin of *Drosophila melanogaster*.** *Genetica* 2000, **109:**9-18.
39. Brosseau GE: **Genetic analysis of the male fertility factors on the Y chromosome of *Drosophila melanogaster*.** *Genetics* 1960, **45:**257-274.
40. Howe M, Dimitri P, Berloco M, Wakimoto BT: ***Cis*-effects of heterochromatin on heterochromatic and euchromatic gene activity in *Drosophila melanogaster*.** *Genetics* 1995, **140:**1033-1045.
41. Koryakov DE, Zhimulev IF, Dimitri P: **Cytogenetic analysis of the third chromosome heterochromatin of *Drosophila melanogaster*.** *Genetics* 2002, **160:**509-517.
42. Devlin RH, Bingham B, Wakimoto BT: **The organization and expression of the *light* gene, a heterochromatic gene of *Drosophila melanogaster*.** *Genetics* 1990, **125:**129-140.
43. Biggs WH, 3rd, Zavitz KH, Dickson B, van der Straten A, Brunner D, Hafen E, Zipursky SL: **The *Drosophila rolled* locus encodes a MAP kinase required in the *sevenless* signal transduction pathway.** *EMBO J* 1994, **13:**1628-1635.
44. Tulin A, Stewart D, Spradling AC: **The *Drosophila* heterochromatic gene encoding poly(ADP-ribose) polymerase (PARP) is required to modulate chromatin structure during development.** *Genes Dev* 2002, **16:**2108-2119.
45. Bonaccorsi S, Gatti M, Pisano C, Lohe A: **Transcription of a satellite DNA on two Y chromosome loops of *Drosophila melanogaster*.** *Chromosoma* 1990, **99:**260-266.
46. Reugels AM, Kurek R, Lammermann U, Bunemann H: **Mega-introns in the dynein gene DhDhc7(Y) on the heterochromatic Y chromosome give rise to the giant threads loops in primary spermatocytes of *Drosophila hydei*.** *Genetics* 2000, **154:**759-769.
47. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, *et al.*: **Finishing a whole-genome shotgun sequence assembly: Release 3 of the *Drosophila* euchromatic sequence.** *Genome Biol* 2002, **3:**research0079.1-0079.14.
48. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Milburn GH, Prochnik SE, *et al.*: **Annotation of the *Drosophila* euchromatic genome: a systematic review.** *Genome Biol* 2002, **3:**research0083.1-0083.22.
49. Dimitri P: **Cytogenetic analysis of the second chromosome heterochromatin of *Drosophila melanogaster*.** *Genetics* 1991, **127:**553-564.
50. Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C: **Splicing signals in *Drosophila*: intron size, information content, and consensus sequences.** *Nucleic Acids Res* 1992, **20:**4255-4262.
51. Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, Harris N, Marshall B, Shu S, Kaminker JS, Prochnik SE, *et al.*: **An integrated computational pipeline and database to support whole-genome sequence annotation.** *Genome Biol* 2002, **3:**research0081.1-0081.11.
52. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer VR, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, *et al.*: **Apollo: A sequence annotation editor.** *Genome Biol* 2002, **3:**research0082.1-0082.14.
53. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8:**967-974.
54. Carvalho AB, Vibranovski MD, Carlson JW, Celniker SE, Hoskins RA, Rubin GM, Sutton GG, Adams MD, Myers EW, Clark AG: **Y chromosome and other heterochromatic sequences of the *Drosophila melanogaster* genome: how far can we go?** *Genetica*, in press.
55. Carvalho AB, Lazzaro BP, Clark AG: **Y chromosomal fertility factors *kl-2* and *kl-3* of *Drosophila melanogaster* encode dynein heavy chain polypeptides.** *Proc Natl Acad Sci USA* 2000, **97:**13239-13244.
56. Carvalho AB, Dobo BA, Vibranovski MD, Clark AG: **Identification of five new genes on the Y chromosome of *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 2001, **98:**13225-13230.
57. Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, *et al.*: **A *Drosophila* full-length cDNA resource.** *Genome Biol* 2002, **3:**research0080.1-0080.8.
58. Tartof KD: **Increasing the multiplicity of ribosomal RNA genes in *Drosophila melanogaster*.** *Science* 1971, **171:**294-297.
59. Williams SM, Robbins LG: **Molecular genetic analysis of *Drosophila* rDNA arrays.** *Trends Genet* 1992, **8:**335-340.
60. Hoskins RA, Nelson CR, Berman BP, Laverty TR, George RA, Ciesiolka L, Naeemuddin M, Arenson AD, Durbin J, David RG, *et al.*: **A BAC-based physical map of the major autosomes of *Drosophila melanogaster*.** *Science* 2000, **287:**2271-2274.
61. Yasuhara J, Marchetti M, Fanti L, Pimpinelli S, Wakimoto BT: **A strategy for mapping the heterochromatin of chromosome 2 of *Drosophila melanogaster*.** *Genetica*, in press.
62. Wakimoto BT, Hearn MG: **The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of *Drosophila melanogaster*.** *Genetics* 1990, **125:**141-154.
63. **RepeatMasker** [http://ftp.genome.washington.edu/RM/RepeatMasker.html]
64. Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle R, George R, Harris N, *et al.*: **An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the *Adh* region.** *Genetics* 1999, **153:**179-219.
65. Lohe AR, Brutlag DL: **Multiplicity of satellite DNA sequences in *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 1986, **83:**696-700.

66. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, *et al.*: **The genome sequence of the malaria mosquito *Anopheles gambiae*.** *Science* 2002, **298:**129-149.
67. **Human Genome Sequencing Center: Baylor College of Medicine** [http://hgsc.bcm.tmc.edu]
68. Reese MG, Kulp D, Tammana H, Haussler D: **Genie - gene finding in *Drosophila melanogaster*.** *Genome Res* 2000, **10:**529-538.
69. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R: **Prediction of probable genes by Fourier analysis of genomic sequences.** *Comput Appl Biosci* 1997, **13:**263-270.
70. Yan CM, Dobie KW, Le HD, Konev AY, Karpen GH: **Efficient recovery of centric heterochromatin *P*-element insertions in *Drosophila melanogaster*.** *Genetics* 2002, **161:**217-229.
71. **Berkeley *Drosophila* Genome Project** [http://www.fruitfly.org]
72. **BACPAC Resources** [http://www.chori.org/bacpac]
73. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic Acids Res* 2002, **30:**2478-2483.
74. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
75. **FlyBase GadFly Genome Annotation Database** [http://www.fruitfly.org/cgi-bin/annot/query]
76. Yamamoto MT, Mitchelson A, Tudor M, O'Hare K, Davies JA, Miklos GL: **Molecular and cytogenetic analysis of the heterochromatin-euchromatin junction region of the *Drosophila melanogaster* X chromosome using cloned DNA sequences.** *Genetics* 1990, **125:**821-832.