Research

Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence

Susan E Celniker*, David A Wheeler†, Brent Kronmiller*, Joseph W Carlson*, Aaron Halpern‡, Sandeep Patel*, Mark Adams‡, Mark Champe*, Shannon P Dugan§, Erwin Frise*, Ann Hodgson§, Reed A George*, Roger A Hoskins*, Todd Laverty¶, Donna M Muzny§, Catherine R Nelson*, Joanne M Pacleb*, Soo Park*, Barret D Pfeiffer*, Stephen Richards*§, Erica J Sodergren§, Robert Svirskas¥, Paul E Tabor§, Kenneth Wan*, Mark Stapleton*, Granger G Sutton‡, Craig Venter‡, George Weinstock§, Steven E Scherer§, Eugene W Myers‡, Richard A Gibbs§ and Gerald M Rubin*¶

Addresses: *Berkeley Drosophila Genome Project, Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. †Human Genome Sequencing Center, and Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA. ‡Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. §Drosophila Sequencing Team, Human Genome Sequencing Center, and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. ¶Howard Hughes Medical Institute, Department of Molecular and Cellular Biology, University of California, Berkeley, CA 94720, USA. ¥Amersham Biosciences, 2100 East Elliot Road, Tempe, AZ 85284, USA.

Correspondence: Susan E Celniker. E-mail:celniker@bdgp.lbl.gov

Published: 23 December 2002

Genome **Biology** 2002, **3(12)**:research0079.1–0079.14

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/research/0079

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

Received: 17 October 2002
Revised: 25 November 2002
Accepted: 27 November 2002

Abstract

Background: The *Drosophila melanogaster* genome was the first metazoan genome to have been sequenced by the whole-genome shotgun (WGS) method. Two issues relating to this achievement were widely debated in the genomics community: how correct is the sequence with respect to base-pair (bp) accuracy and frequency of assembly errors? And, how difficult is it to bring a WGS sequence to the accepted standard for finished sequence? We are now in a position to answer these questions.

Results: Our finishing process was designed to close gaps, improve sequence quality and validate the assembly. Sequence traces derived from the WGS and draft sequencing of individual bacterial artificial chromosomes (BACs) were assembled into BAC-sized segments. These segments were brought to high quality, and then joined to constitute the sequence of each chromosome arm. Overall assembly was verified by comparison to a physical map of fingerprinted BAC clones. In the current version of the 116.9 Mb euchromatic genome, called Release 3, the six euchromatic chromosome arms are represented by 13 scaffolds with a total of 37 sequence gaps. We compared Release 3 to Release 2; in autosomal regions of unique sequence, the error rate of Release 2 was one in 20,000 bp.

Conclusions: The WGS strategy can efficiently produce a high-quality sequence of a metazoan genome while generating the reagents required for sequence finishing. However, the initial method of repeat assembly was flawed. The sequence we report here, Release 3, is a reliable resource for molecular genetic experimentation and computational analysis.

Background

The genome of *Drosophila melanogaster* was sequenced using a whole-genome shotgun (WGS) approach [1,2]. The first assembly (WGS1) used only plasmid and BAC paired-end sequences. The second added BAC and P1-based finished and draft sequences (see Table 3 of [1] for details); these two assemblies were compared in [1]. The joint assembly was submitted to GenBank as Release 1. This sequence contained many gaps and regions of low sequence quality. A second release, Release 2, corrected some errors in the order and orientation of small scaffolds present in Release 1, and filled a few hundred very small sequence gaps. Using improved WGS sequence-assembly algorithms, two additional assemblies of just the WGS plasmid and BAC paired end sequences used in WGS1 were generated in March 2001 (WGS2) and July 2002 (WGS3), roughly coinciding with the WGS assemblies of the human [3] and mouse genomes [4], respectively.

This paper describes the finishing work we have done to improve Release 2 in order to generate the Release 3 sequence of the *D. melanogaster* euchromatin. The status of the heterochromatic regions of the genome is reported in [5]. Our goals in generating the euchromatic portion of Release 3 were to close all the gaps, improve regions of low sequence quality, extend the sequence at the telomeric and centromeric ends of each chromosome, and verify the whole genome assembly. The Release 3 euchromatic genome sequence has been reannotated [6,7] using a new annotation tool, Apollo [8], and deposited in GenBank. Companion papers [7,9] address the complete reannotation of the *Drosophila* euchromatic genome on the basis of improved genomic sequence and new expressed sequence tag (EST) and cDNA sequences. The improvements made to the genomic sequence in Release 3 had a large impact on the annotation of transposable elements [9] because of the substantial corrections made in the assembly of repeated sequences. Because the non-repetitive regions of the genome were generally of good quality in Release 2, most of the improvements to the annotation of these regions resulted from the increased amounts of EST and cDNA data [7]. Release 3 provides a euchromatic sequence that is virtually gap free and of high accuracy. We were thus able to rigorously assess the quality of Release 2, as well as the sequences generated by each of the three WGS assemblies, by simply comparing these sequences to the Release 3 euchromatic sequence and assuming that all the differences were the result of errors or omissions in the other sequence.

Results and discussion

Overview of the finishing strategy

Both of the sequencing centers that participated in the finishing effort, Lawrence Berkeley National Laboratory (LBNL) and the Human Genome Sequencing Center, Baylor College of Medicine (HGSC), have extensive experience in finishing the sequence of individual BAC clones. To take advantage of this expertise, we reduced the problem of finishing a WGS sequence to one of finishing a set of overlapping BAC clones. This approach allowed us to use the sequence assembly software phrap [10], which provided an independent assembly of the raw trace data and an estimated error rate. Thus, apart from filling gaps and improving sequence quality, we were able to compare two independent assemblies of the same trace data. LBNL took responsibility for chromosome arms 2L, 2R, 3R, 4 and the proximal half of the X chromosome (80.6 Mb), and HGSC chromosome arm 3L and the distal half of the X chromosome (36.4 Mb).

We previously generated a physical map of the major autosomes [11]. BAC-based maps of the X chromosome [12] and chromosome 4 [13] were also available (Figure 1). We selected a tiling path of BAC clones spanning the physical maps and Release 2 sequence using the available sequence-tagged site (STS) maps and BAC end sequences. The WGS sequence traces and end sequences of the tiling path BACs were assigned coordinates based on their order in the Release 2 assembly. We then binned all the traces belonging within a particular BAC and extending 500 bp beyond the end sequence of that BAC. These traces were combined with sequence traces (1.5x sequence coverage) that had been generated from libraries made from the individual tiling-path BACs [1] and assembled using phrap. These assemblies formed the starting point for our finishing efforts.

To generate Release 3, we closed gaps by sequencing 518 3-kb subclones, 565 10-kb clones and 290 fragments generated by PCR; we improved quality with 15,344 custom primer sequence runs. To assess the accuracy of the individual BAC assemblies, we compared *in silico* restriction maps generated from the sequence of each BAC to restriction fingerprints essentially as described [14]. A set of *Bam*HI, *Eco*RI and *Hind*III digests was used to verify the assembly of 618 BACs. For 328 BACs, only two of the three enzyme digests were available to verify the assembly. The finished BAC clone sequences were submitted individually to GenBank.

The assembly of the individual BACs into entire chromosome arms was verified in two ways. First, BAC tiling path clones were mapped by *in situ* hybridization to the polytene salivary gland chromosomes of third instar larvae, which serves as an unambiguously correct physical map. *In situ* hybridization data for 547 clones in the BAC tiling path have been reported previously [11]. Images documenting the localization of 915 (96%) BACs in the tiling path can be found in the GadFly Genome Annotation Database [15]. Second, the unique end sequences of 8,424 BACs were aligned to the assembled chromosome arm sequences. If the fully sequenced BACs were not assembled in the correct order and orientation, we would expect to find regions of the genome where those end sequences fail to align to positions between 100 and 200 kb apart; such regions of misaligned paired BAC end sequence

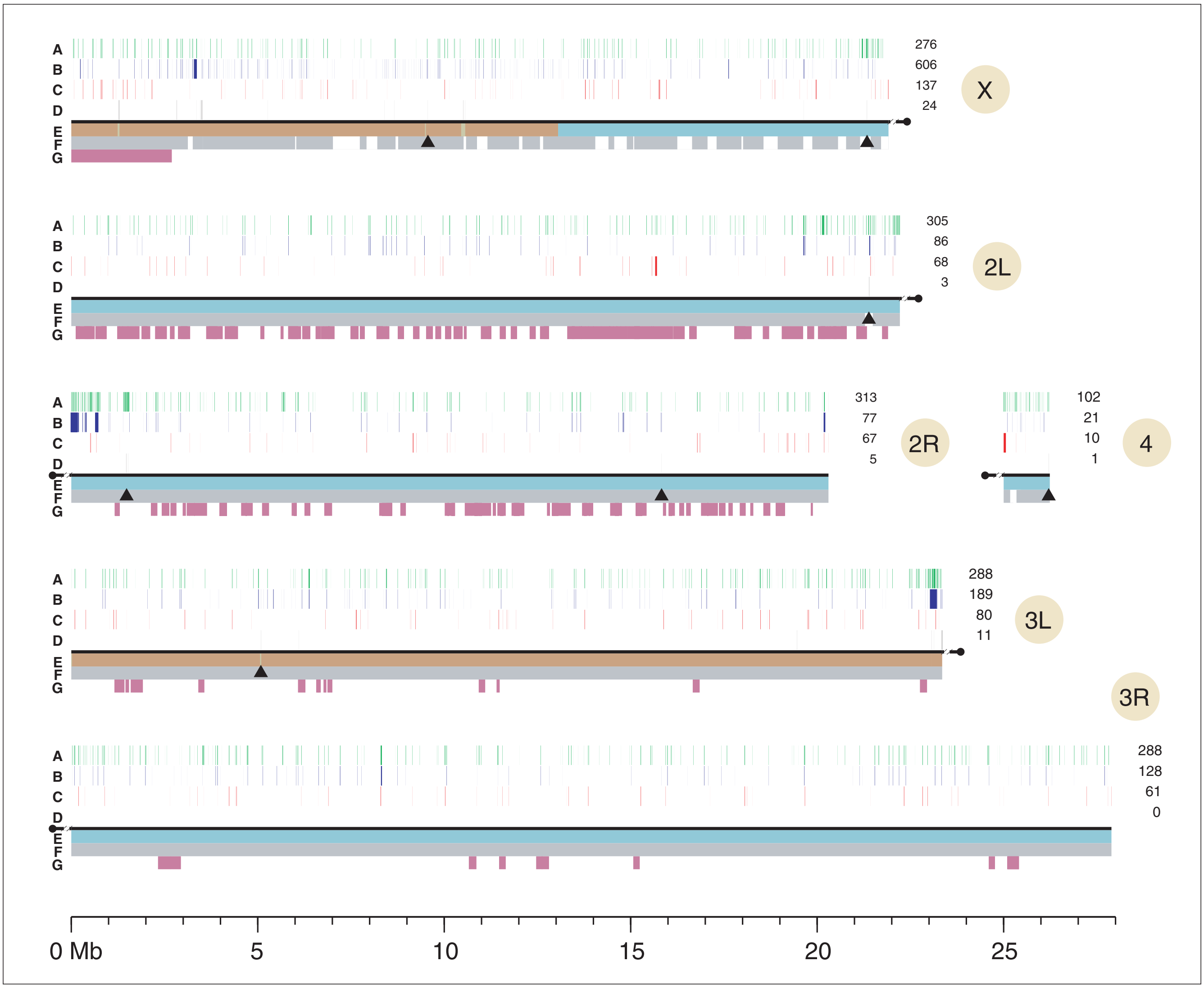

Figure 1
Status of the *Drosophila melanogaster* euchromatic genome. Each chromosome arm is represented by a black horizontal line with a circle indicating its centromere. For each arm, seven tiers of information (A-G) are presented. (A) Each vertical green line represents the position of a transposable element. (B) Each vertical blue line represents the position of a 'declared' gap in Release 2. (C) Each vertical red line represents the position of an 'undeclared' gap in Release 2 greater than 20 bp, detected by comparing the Release 2 and Release 3 sequences. (D) Each vertical black line represents the position of a sequence gap that remains in Release 3. (E) The horizontal bars depict the regions of the genome assigned to LBNL (blue) or the HGSC, Baylor College of Medicine (brown) for generating Release 3. (F) The gray horizontal bar represents the status of the physical maps that supplied the initial BAC tiling paths for sequencing; presence of the gray bar indicates an available BAC contig. The sources of these BAC maps were as follows: chromosome X [12,50], chromosome arms 2L, 2R, 3L, and 3R [11] and chromosome 4 [13]. The black triangles represent the seven physical map gaps remaining in the euchromatic portion of the genome in Release 3. (G) The purple bar represents the position of cosmid, P1 or BAC clones that had been completely sequenced prior to Release 2. Those at the telomere of chromosome X were sequenced by the EDGP [51]; the other clones were sequenced by the BDGP at LBNL [1]. The numbers to the left of rows A, B, C and D are the chromosome arm totals for each category plotted. The scale in million bases (Mb) is shown at the bottom of the figure.

were not observed in the final Release 3 chromosome arm assemblies. The Release 3 chromosome arms are composed of 13 sequence scaffolds and contain 31 sequence gaps. A sequence scaffold is defined as a set of contiguous sequence contigs, ordered and oriented with respect to one another; within a scaffold, the gaps between adjacent contigs are of known size and are spanned by clones [1]. Physical map gaps are gaps between scaffolds; in these cases, no clones that span the gap have been identified.

Refinements to the physical map

A BAC-based physical map of the X chromosome was constructed by the European Drosophila Genome Project (EDGP) [12] and used as a starting point for BAC-based sequence assembly of the X. In our finishing work, we replaced mismapped BACs at 13A and 14D. We also filled nine clone gaps in the physical map, identifying BACs that span them by comparing the paired end sequences of 9,869 BACs [1] with the Release 2 sequence assembly. In Release 3, the X chromosome is in three scaffolds of 10.5, 10.8 and 0.4 Mb, and has two physical map gaps at 9EF and 20A, estimated to be 150 kb and 200 kb in size, respectively (Table 1). Several short tandem arrays lie within 3 kb of the 9EF gap, but it possesses no other remarkable sequence features. Complex nests of transposable elements flank the gap at 20A, which lies within the centric heterochromatin; such nests of transposable elements are common in the proximal regions of the chromosome arms [9].

Our BAC-based physical map of the autosomes [11] was used as a starting point for sequence assembly of the second chromosome. The left arm of chromosome 2 (2L) is in two scaffolds of 21.7 and 0.8 Mb and has one clone gap at 39D-E that contains 100-200 tandem repeats of the histone gene cluster [16]. Each repeat contains five histone genes: *Histone H1 (His1), Histone H3 (His3), Histone H4 (His4)*, *Histone H2A (His2A)* and *Histone H2B (His2B)*. Two predominant forms of the repeat, 4.8 and 5.0 kb, have been previously described and differ primarily in the size of the spacer DNA between *His1* and *His3*. The variants are not interspersed with one another; but form segregated clusters. The BAC at the distal side of the gap has at least three copies of the 5-kb variant and the BAC at the proximal side of the gap has at least two copies of the 4.8-kb variant.

The right arm of chromosome 2 (2R) is in three scaffolds of 1.5 Mb, 14.3 Mb and 4.5 Mb, with two clone gaps, one in 42B and another in 57B, estimated to be 100 kb and 300 kb, respectively (Table 1). The clone gap in 42B is associated with 50 to 100 copies of a previously uncharacterized 596 bp repeat showing weak similarity to RNA-directed RNA polymerase, interspersed with 1-kb units that have 49% nucleotide similarity to the *1731* transposable element. The other clone gap on 2R, in 57B, is flanked on the distal side by 120-bp tandem repeats with similarity to snoRNAs [17], and on the proximal side by a different 120-bp repeat.

The left arm of chromosome 3 (3L) is in two scaffolds of 5 Mb and 18.5 Mb, with a single clone gap at 64C, estimated to be 100 kb (Table 1). The right arm of chromosome 3 (3R) has no euchromatic clone gaps and is represented in a single sequence contig of 27.9 Mb.

A BAC-based physical map of chromosome 4 [13] was used as a starting point for sequence assembly of this arm. This map has two clone gaps, one at 102A and the other at 102EF. The gap at 102A was estimated to be very small and was filled with 46,766 bp of sequence in Release 3. The gap at 102EF was estimated to be 100 kb and remains the only clone gap on chromosome 4 in Release 3. In Release 3, chromosome 4 exists in two scaffolds 1.2 Mb and 0.03 Mb. The gene *CG17467* extends into the gap from the proximal side, and the *CG18026 (CAPS)* gene extends into the gap from the distal side; these sequences can be used as starting points to fill the gap for Release 4 of the genome sequence. The *CAPS* gene is found in a 64-kb scaffold in WGS3 that extends 28.5 kb into the gap, ending in approximately 2 kb of the simple 9-bp repeat CATAATAAT.

Assessment of the quality of Release 3

The total size of the euchromatic portion of the *Drosophila* genome in Release 3 is 116,914,271 bp. The status of each chromosome arm, including length, number of physical map gaps, number of finished and unfinished BACs, number of sequence gaps, and sequence quality, estimated as error rate in a sliding window of 100 kb, is shown in Table 1. The positions of the remaining gaps are diagrammed in Figure 1. The euchromatic genome is assembled into 50 contigs; 50% of the genome is contained in contigs (N50) of 14 Mb or greater, and 99% (N99) is in contigs greater than 526 kb. A total of seven physical map gaps and 37 sequence gaps remain (Table 1). The sequence is highly accurate: 98.7% of the base pairs are contained within 100-kb regions having estimated error rates of less than one per 100,000 bp. The estimated error rate was also determined for each of the 950 BACs that comprise the tiling path. All have a phrap estimated error rate of less than one in 30,000 bp, and 875 (92%) are estimated to have less than one error in 100,000 bp.

Extending the sequence

Our finishing efforts made modest extensions toward the telomere and into the euchromatin-heterochromatin boundary region at the base of each chromosome arm. We can recognize the telomeres by their distinctive sequence features: variable numbers of *HeT-A* and *TART* transposable elements at their extreme ends, followed proximally by variable numbers and types of Taq minisatellite repeats (TAS elements [18]). The transition from euchromatin to heterochromatin is gradual. The boundary between euchromatin and centric heterochromatin has been defined cytogenetically [19], but at the molecular level it is simply characterized by a significant increase in the density of transposable elements [9] and other repetitive DNA sequences. We report in this paper on that portion of the genome contained in the mapped scaffolds of Release 2, with the exception of four small scaffolds on chromosome 2 that we know lie in heterochromatin; these regions include all of the euchromatin and extend into heterochromatin, as defined cytogenetically [5].

The sequences found at the telomeric and centromere proximal ends of each chromosome arm in Release 3 are as follows.

Chromosome X

The sequence at the telomere of the X chromosome in Release 3 is contained in BACR13J02, but this BAC is not completely

Table 1

Status of Release 3

			Physical map gaps							Estimated error rate*		
Chromosomal region	Group	Size	Number	Location	Estimated maximum size†	Finished BACs	Unfinished BACs	Sequence gaps‡	Release 2 sequence	10^4 to 10^5	10^5 to 10^6	>10^6
X (1-11)	HGSC	13,053,575	1	9EF	150 kb	85	14§	22	234,520¶	2	16	241
X (12-20)	LBNL	8,921,907	1	20B2	200 kb	73¥	2#	2	0	19	76	84
2L	LBNL	22,217,931	1	39D	~500kb - 1 Mb	177	2**	3	0	14	30	397
2R	LBNL	20,302,755	2	42B 57B	100 kb 300 kb	159	4††	5	0	11	56	335
3L	HGSC	23,352,213	1	64C	100 kb	175	9‡‡	11	47,653§§	6	50	409
3R	LBNL	27,890,790	0	NA	NA	235	0	0	0	8	119	430
4	LBNL	1,237,864	1	102F	100 kb	14	0	1	0	3	7	13
Total		116,914,271	7			917	31	44		63	354	1,909

*Estimated error rates were determined for 100-kb bins, chosen to overlap by 50 kb. Estimated error rates were determined for bins containing sequence or physical map gaps. However, gaps represented by Ns in the sequence did not contribute to the estimated error rate; thus, the error rate reflects only those sequences present. †*In situ* hybridization of flanking clones to polytene chromosomes and estimates of DNA content per band [47] allowed us to estimate the maximum size of the clone gaps. All of the gaps are in regions of tandem repeats and the flanking BACs extending into the gap might contain sufficient amounts of the repeat to lead to a misleading *in situ* mapping result. Therefore, we also examined the next BAC in the tiling path, not containing the repeat, to ensure we were using a unique sequence probe. Four BACs are listed for each gap, two on each side, in the order they occur in the genome. The gap at 9EF is flanked by BACR48E06 (location, 9C2-E1), BACR10I17 (ND) and BACR26N01 (9F1-10A2), BACR17B23 (10A1-2). The gap at 20B2 is flanked by BACR23I18 (19F3-A2) BACR22O16 (20A3-B2) and BACR06L03 (20B2-C2), BACR05K22 (20C1-2). The gap at 39D was sized by estimating the histone repeat copy number [16]. The estimate from the flanking BACs, BACR34H23 (39A6-C3) and BACR03L08 (39F1-2) is 400 kb. The gap at 42B is flanked by BACR13P06 (42A3-19), BACR36A03 (42B1-2) and BACR28N07 (42B1-3), BACR01C10 (42B3-C6). The gap at 57B is flanked by BACR03N16 (57A1-4), BACR08P05 (57A5-B3), and BACR10P16 (cytology 57B2-6), BACR04E05 (57B4-6). The gap at 64C is flanked by BACR23H09 (64B15-C2), BACR17L24 (64C1-4), and BACR12G07 (64C5-12), BACR12P14 (64C9-12). The gap at 102F is flanked by BACR13D24 (102D6-E6), BACR22J20 (102E3-F2), and BACH59K20 (102F1-5), BACN05O16 (cross-hybridizes to all telomeres, consistent with its location at the chromosome end). ‡This number includes all instances where we inserted a string of Ns to indicate missing sequences; it is the sum of physical map gaps and gaps due to failure to complete the sequence of cloned regions. In some cases a single physical map gap results in more than one sequence gap. For example, all three sequence gaps on 2L are found in the unfinished BACs that extend into the histone repeat region and four of the five sequence gaps on 2R are found in the unfinished BACs that extend into the repeat region of 42B. Excluding the physical map gaps, the gaps on X 1-11 total 60.6 kb; the gaps on 2R total 1,549 bp; the gaps on 3L total 26.2 kb, excluding the two gaps mapping to heterochromatin. There are no gaps, other than physical map gaps, on 2L, 4 or X 12-20. §The Release 3 sequence of chromosome X 1-11 includes sequence from 14 unfinished BAC clones. Each of these BACs contains one or two regions of repeat sequence that are difficult to resolve. Eight of the unfinished clones contain *Foldback* (BACR40O10, BAC23M02, BACR19G09, BACR26B05, BACR29A04), multiple or rearranged *roo* (BACR17E02, BACR46E23) or *412* (BACR07P13) elements. Six of the clones (BACR01A14, BACR17E02, BACR19D19, BACR25I09, BACR29B18, BACR39C15) contain duplications of other, uncharacterized, repeats. BACR13J02 is the most distal clone in Release 3, extending the Release 2 assembly by approximately 15 kb. Seven of the 14 BACs that were unfinished at the time of Release 3 have since been finished. Five clones (CHORI 22340I08, BACR32E02, CHORI 221-14P20, CHORI 221-17A11 and CHORI 223-05O10) have been added to the tiling path to span the genomic regions that are still represented by Release 2 sequences (see ¶); these BACs were not sequenced for Release 3. 366 bp of sequence (coordinate 3.4 Mb, cytology 3EF) are not contained within a BAC but are spanned by 10-kb genomic clones. The EDGP identified two clones, BACR37M19 and BACR20K04, as mapping to this region [12] but we determined that their end sequences align elsewhere. The BAC clone coverage of the X chromosome is expected to be lower than the BAC clone coverage of the autosomes and may explain the BAC clone gap in 3EF. BACs whose names begin with CHORI are derived from a library made with randomly sheared DNA [48]. ¶Four Release 2 segments not covered in finished BACs were used to produce the Release 3 sequence (see Materials and methods, Arm assembly and overlap verification): 18.3 kb starting at position 1,262,967 bp; 104 kb starting at position 3,412,482 bp; 12.2 kb starting at position 9,489,057 bp; 99.7 kb at starting at position 10,462,912 bp. The latter segment extends into the clone gap at 9EF. ¥The last 36 kb of sequence at the centromeric end of the X chromosome are not contained within a BAC and are derived from a phrap assembly using WGS traces and the complete sequence of two 10-kb genomic clones. #One of the two unfinished BACs (BACR22O16) extends into the physical map gap and the second (BACR39I01) contains a sequence gap resulting from our inability to assemble a difficult repetitive region that includes at least eight copies of a 4.7 kb tandem repeat having similarity to a degenerate *mdg3* transposable element lacking LTRs. **These two unfinished BACs (BACR05D08 and BACR43O11) flank and extend into the 1-Mb histone gene cluster. ††Three unfinished BACs (BACR48D05, BACR03A06 and BACR36A03) extend into the gap at 42B and one unfinished BAC (BACR08P05) extends into the 57B gap. ‡‡The nine unfinished BACs are BACR31B14, BACR43N11, BACR27G13, BACR29O22, BACR01D04, BACR01B21, BACR09G21, BACR30I05 and BACR34K23. BACR31B14, BACR43N11, and BACR27G13 contain sequence gaps that are a consequence of transposable elements (*FB* or *roo*) with complex internal rearrangements, tandem repeats or deletions. Two BACs, BACR29O22 and BACR01B21, contain a *roo* and a *Doc* element, respectively, and were not completed. One sequence gap in BACR01D04 is the result of a small misassembly that could not be resolved. Three other sequence gaps are in an unfinished segment of clone BACR34K23. Three (BACR09G21, BACR30I05 and BACR27G13) of the nine BACs that were unfinished at the time of Release 3 are now finished. Five clones (BACR29A07, CHORI 223-12D09, BACR15L14, CHORI221-06A19 and BACR03B05) have been have been added to the tiling path to span the genomic regions that are still represented by Release 2 sequences (see §§); these BACs were not sequenced for Release 3. The addition of BACR29A07 to the tiling path corrects an inversion in Release 3 at the 3L centromere. The BAC order is now BACR17M18, BACR29A07, BACR22B15 and BACR34K23. In addition, there are 13 finished BACs from 3L that have been submitted to GenBank with unresolved tandem repeat annotations, in accordance with the G16 finishing standards for the human genome project [49]. §§Three Release 2 segments not covered in finished BACs were used to produce the Release 3 sequence (see Materials and methods, Arm assembly and overlap verification): 10.8 kb starting at position 1, 18.9 kb starting at position 5,065,167, 12.6 kb starting at position 23,339,636 bp. The 18.9 kb sequence extends into the 64C clone gap. The 12.6 kb sequence contains two gaps mapping to BACR30H12.

assembled and does not extend to the TAS repeats. The Release 3 sequence of the X chromosome extends proximally approximately 10 kb beyond Release 2 toward the centromere. The proximal 38 kb of Release 3 are not covered by BAC clones; we used 10-kb clones from the WGS in order to produce this sequence. The WGS3 assembly extends beyond Release 3 by 42 kb, ending in 235 copies of the simple repeat TAGA, a known heterochromatic satellite repeat [20].

Chromosome 2L

The Release 3 telomeric sequence, extending approximately 5 kb farther than that of Release 2, ends in 11 copies of the TAS repeat. The proximal sequence of chromosome arm 2L ends in 15 kb of nested transposable elements.

Chromosome 2R

The Release 3 telomeric sequence ends in two copies of a 235-bp repeat similar to the previously described telomeric sequence in the TAS element. The proximal sequence of chromosome arm 2R is highly enriched in transposable element sequence.

Chromosome 3L

There has been no change in the sequence of the 3L telomere in Release 3 compared to Release 2. It has a single copy of a 458-bp sequence with 98% similarity to the minisatellite TAS repeats of 2L. The proximal sequence of chromosome arm 3L extends beyond Release 2 by 90 kb and is highly enriched in transposable element sequence.

Chromosome 3R

The telomere of chromosome arm 3R in Release 3 contains six copies of a 983-bp repeat similar to the previously described telomeric sequence in the TAS element; it extends the Release 2 sequence by approximately 6 kb. We did not substantially extend the proximal sequence of chromosome arm 3R in Release 3. The WGS3 assembly extends 774 bp beyond Release 3 and ends in 67 tandem copies of a nearly perfect simple repeat TATAA, a known heterochromatic satellite repeat.

Chromosome 4

The most distal BAC on chromosome 4, BACH59K20, is not anchored to the 1.2-Mb main scaffold. This BAC contains over 32-kb of sequence not mapped to chromosome 4 in Release 2. Analysis of the sequence of this BAC shows that it should have been appended in inverted orientation, because it ends in two copies of an approximately 700 bp portion of a *HeT-A* element, whose poly(A) tail is oriented toward the centromere. The sequence of chromosome arm 4 in Release 3 extends proximally to that of Release 2 by nearly 60 kb. This extension contains four 8 kb tandem imperfect repeats, each bounded by a *narep 1* repeat. The distal repeat contains the *plexin B* gene; the three proximal repeats contain related protein-coding genes *CG32010*, *CG32011* and *CG32009*.

Comparing Releases 2 and 3

The Release 2 sequence of the euchromatin contained 1,107 gaps of which we were aware. These 'declared' gaps were represented in the sequence by a string of Ns corresponding in length to the estimated gap size, or by 1,000 Ns if we were unable to estimate the size of the gap. Our comparisons of Releases 2 and 3 identified 424 additional insertions or deletions greater than 20 bp, which we refer to as 'undeclared' sequence gaps (Table 2, Figure 1). Now that more than 97% of the gaps in the euchromatic sequence have been filled, we

Table 2

Sequence content of gaps in Release 2

	X		2L		2R		3L		3R		4		Subtotals		Total
Total gaps	743		154		145		269		189		31				1531
	D*	U†	D	U	D	U	D	U	D	U	D	U	D	U	
	606	137	86	68	77	68	189	80	128	61	21	10	1107	424	1531
															0
Content															
TEs	61	42	48	33	42	42	52	47	50	39	9	5	262	208	470
Simple repeats	353	10	19	9	15	4	109	2	38	9	10	2	544	36	580
Homopolymers	10	0	1	0	3	0	2	0	3	0	0	0	19	0	19
Unique sequence	150	23	13	8	8	11	21	24	28	3	1	0	221	69	290
Tandem repeats	14	34	1	12	3	6	0	0	1	10	0	1	19	63	82
Missassemblies	3	18	1	4	1	2	0	1	7	0	1	0	13	25	38
Gross misassemblies	1	0	1	1	3	0	0	0	1	0	0	1	6	2	8
Not yet determined	14	10	2	1	2	3	5	6	0	0	0	1	23	21	44

Analysis of the sequence gaps in Release 2 determined by comparison with Release 3 (see text for details).*Declared (D) gaps represented in Release 2 by sets of Ns. †Undeclared (U) gaps not recognized in Release 2, and identified by comparison to Release 3.

are able to examine the sequences that were missing in both the declared and undeclared gaps. The sequences missing in each gap were classified into seven categories: unique sequence, transposable element, homopolymer, simple repeat, tandem duplications, local misassembly, or gross misassembly. In addition, we identified individual base-pair differences, as well as insertions and deletions of less than 20 bp.

In the WGS assemblies, the sequence coverage of the X and Y chromosomes is expected to be less than for the autosomes. The autosomal coverage was estimated to be 12x [1]; thus, sequence coverage is expected to be 9x for the X chromosome, and 3x for the Y, assuming an equal number of male and female embryos in the collection used to make the DNA for the WGS. Presumably as a result of its reduced sequence coverage, the Release 2 sequence of the X chromosome has 2.5 times as many gaps (743, representing 49% of the total) per million bases as those of the autosomes and more than half the gaps in the euchromatic portion of the genome that lie in unique sequence map to the X chromosome. The Y chromosome is almost entirely heterochromatic and Y chromosome sequences are limited to small WGS scaffolds [21].

Repetitive elements are difficult to assemble in a whole-genome shotgun strategy. As a consequence, one of the initial steps in the Release 2 assembly was to identify known repetitive elements, including transposable elements, and remove their traces from the early steps of the assembly process [2]. After the unique sequence was assembled and sequence contigs were ordered and oriented, those transposable element sequence traces were added back. Sequence traces belonging to transposable element families were assembled using an aggressive strategy that resulted in composite sequences for most of the transposable elements in Release 2. Part of the finishing effort has been to determine the sequence of each individual element. Of the 1,572 full and partial transposable elements in Release 3 [9], only 380 are identical to the corresponding Release 2 transposable element sequences. Of these identical elements, 323 are shorter than 900 bp and could be assembled with two reads if those reads contained sufficient unique sequence to be unambiguously placed. The sequences of the remaining 57 transposable elements are identical between Release 2 and Release 3 because their sequence was determined by the Berkeley Drosophila Genome Project (BDGP) as part of the finishing process leading to Release 2 [1]. Approximately a third of the total number of gaps in Release 2 euchromatin are the consequence of transposable elements. A slightly higher fraction of the gaps are the consequence of simple repeats.

Duplications in the *Drosophila* genome are rare in comparison to the number found in mammalian genomes. We observed tandem repeats whose repeat units range in size from 10 bp to 30 kb. The largest region comprised of tandemly duplicated repeats in the euchromatic portion of the *Drosophila* genome is the histone cluster at 39B described above. As a prelude to a more rigorous analysis of repeats, we have searched for perfect direct or inverted repeats over 100 bp in length that are not transposable elements or low-complexity sequence. We identified 0.9 Mb of repeated sequence (less than 1% of the euchromatic genome), 40% mapping to chromosome X, 40% to chromosome 2, 15% to chromosome 3 and 3% to chromosome 4. Although there are more repeats on chromosomes X and 2, they appear to be randomly distributed and are not clustered at the centromeres or telomeres. This is in contrast to the increased number of tandem duplications found in the pericentric regions of mammalian chromosomes [22]. Misassemblies of tandem repeats account for only 5% of all gaps. Local misassemblies resulting in small insertions, the result of low-quality sequence in Release 2, are also rare, constituting only 2% of the gaps.

We identified only seven gross misassemblies in Release 2. Six resulted in large sequence insertions or deletions in Release 2, and one resulted in a large inversion. In addition, a polymorphism in the isogenic strain resulted in an eighth large-scale sequence difference between Releases 2 and 3. This polymorphic tandem duplication consists of more than 30 kb of sequence bounded by *hobo* and *cruiser* transposable elements; half of the BACs mapping to this location contain one copy of the repeat and the other half have two. Transposable element sequence is associated with the ends of all of the misassembled regions. On the X chromosome, at position 3.4 Mb, 88 kb of complex transposable element sequence in Release 2 has been replaced in Release 3 with a single copy of a *roo* element and a gap. When complete, this region is expected to contain three *roo* elements. On 2L there is one misassembly, at 1.8 Mb. A declared gap estimated to be 1,400 bp and involving an *mdg3* transposable element was filled in Release 3 with 50 kb of sequence. On 2R there are three misassemblies, mapping to 0.8 Mb, 14.8 Mb and 20.1 Mb. The first is associated with four declared gaps within 135 kb of Release 2 sequence. That sequence has been replaced with 60 kb of sequence in Release 3, bounded by a *G* element at one end and an *invader 3* element at the other. The second misassembly is a declared gap, estimated to be 11 kb, that was filled with 37 kb of sequence in Release 3. In the third, 54 kb of complex transposable element sequence in Release 2 has been replaced with a single complete 10-kb *roo* element in Release 3. On 3R, a 6-kb segment in Release 2 is replaced in Release 3 with 33 kb. The Release 2 sequence of chromosome 4 contains a large misassembly that inverts the distal third of the chromosome (John Locke, personal communication). The inversion occurred at position 751,419, the location of a portion of a *1360* element. The misassembled Release 2 sequence ends with a $\mathrm{TAATAATA}_{(27)}$ repeat that maps starting at position 817,409 bp in Release 3.

We identified single base substitutions and single nucleotide insertions and deletions between Release 2 and Release 3.

Excepting transposable elements, the rate of base changes is one per 22,000 bp for the autosomes and one per 5,000 bp for the X chromosome. Within transposable elements, the rate of base changes is one per 124 bp; this high error rate is largely a consequence of comparing the Release 3 sequence of individual elements to the composite sequences in Release 2.

Comparison of WGS assemblies to Release 3

In addition to comparing the Release 3 sequence to the two sequence releases previously deposited in GenBank, we also compared Release 3 to each of three pure WGS assemblies. These comparisons allow an assessment of how well a WGS approach works on a 180 Mb metazoan genome. Although the specific software we used is proprietary, the underlying algorithms and the logic have been fully described [2,3] and the basic approach replicated by others [23,24]. An analysis of these assemblies in light of the modifications to the algorithmic strategy made between assemblies should inform the development of WGS assembly software in general. The sequence traces used in these assemblies are available from the GenBank trace repository and Release 3 offers a high-quality sequence for comparison, providing an important resource for testing new WGS assembly algorithms. The sequences that comprise the WGS2 and WGS3 assemblies are also available for benchmarking.

Three assemblies of the WGS data were made, each using a different version of the Celera Genomics assembly software. WGS1 used 3,156,000 paired-end shotgun reads from genomic plasmids and 12,152 paired BAC-end reads. An additional 62,000 reads that had been erroneously removed during the quality screening process for WGS1 were added to the input data set for the next two assemblies (WGS2 and WGS3). The assembly algorithm that produced WGS1 did not adequately handle repeat resolution; a variety of improvements, including a different method of placing repeat traces, better identification of repeat traces, and the introduction of a sequence correction algorithm, resulted in better subsequent assemblies, particularly of repetitive regions.

The output of the assembly process is a collection of scaffolds made up of ordered and oriented contigs linked together by paired end sequences. Paired-end reads of clones whose size falls within a tight distribution allow the length of every gap between two contigs of a scaffold to be characterized with an expected mean and standard deviation. There were 732,000 pairs of sequences produced from the ends of 2-kb plasmid libraries, another 548,000 pairs of sequences from the ends of 10-kb plasmid libraries, and 12,152 pairs of BAC-end sequences a mean distance of 130 kb apart from each other. Table 3 presents statistics characterizing the scaffolds in the three assemblies. Table 4 presents the same statistics for the scaffolds that align to the Release 3 euchromatic sequences. The scaffolds that do not align to the chromosome arms, totaling 16.5 Mb of sequence and spanning 20.7 Mb, derive from the heterochromatic regions of the genome (see [5] for more information on the content of these scaffolds.)

Table 3

Scaffold, contig and gap statistics for the three assemblies

	WGS1	WGS2	WGS3
Number of scaffolds	816	2,198	2,775
Total Mb spanned	122.92	133.47	137.6
Total Mb of sequence	119.52	129.12	132.94
N50 scaffold length (Mb)	10.70	14.26	13.68
Number of gaps	2,926	5,319	4,936
Number of intra-scaffold gaps	2,110	3,121	2,161
Mean contig length (kb)	40.8	24.3	26.9
Mean gap length (bp)	1,611	1,395	2,190

Table 4

WGS scaffolds that align to the euchromatic portion of Release 3

	WGS1	WGS2	WGS3	Release 3
Number of scaffolds covering Release 3	55	63	53	13
Total Mb spanned	116.39	117.44	117.6	116.91
Total Mb of Release 3 spanned	116.4	116.5	116.8	-
Total Mb of sequence	114.15	115.83	116.42	116.87
Total Mb of Release 3 sequence	114.1	115	115.6	-
N50 scaffold length (in Mb)	10.85	14.45	13.89	18.5
Number of gaps	2,173	2,315	1,130	44
Mean contig length (kb)	52.2	49.5	102	2,335
Mean gap length (bp)	1,531	912	1,335	-

The output of the assembler contains only the sequence reads that assemble with high confidence into a contig. As a result, 15-25% of the sequencing reads, most of which come from heterochromatin or interspersed repetitive elements, do not form part of an assembly. Table 3 provides evidence that the total amount of data that was reliably assembled increased as the assembly algorithms improved. In the later assemblies, more sequence from heterochromatic regions of the genome is reported, but in a more fractured state than for euchromatic regions. These additional assembled pieces, although constituting a modest fraction of the sequence, constitute a large percentage of the number of scaffolds and contigs. The 10 largest scaffolds of the three assemblies constitute over 80% of the sequence and are almost identical.

In all three assemblies, a small number of large scaffolds cover the 116.9 Mb of the Release 3 euchromatic sequence (Table 4). WGS3 scaffolds span 99.91% of Release 3 euchromatic sequence and extend beyond it by 0.84 Mb. Mean gap length increased between WGS2 and WGS3, but there are half as many gaps, indicating that WGS3 resolved many of the smaller interspersed repeats. Scaffold N50 lengths are similar in each assembly (Table 4), as all of the mapped scaffolds are large. Of the 116.9 Mb of Release 3 euchromatic sequence, WGS1 provided 97.6%, WGS2 provided 98.4%, and WGS3 provided 98.9% of the sequence.

The content of sequence gaps in WGS1, WGS2 and WGS3 was determined by comparison to Release 3. The majority of this sequence - 73%, 64%, and 70%, respectively - is repetitive, either tandem repeats or transposable elements.

We compared the order of the Release 3 euchromatic sequences with the order of the contigs and scaffolds in each of the WGS assemblies (see Materials and methods). Assuming that Release 3 is correct, we manually examined and categorized these discrepancies. Table 5 shows the number of incorrectly ordered segments in each category and the total number of base pairs involved in those segments. Occasionally, what look like separate scaffolds actually overlap; a contig at the end of one scaffold contains sequence that belongs in a gap at the end of the other scaffold. These interleaved scaffolds induce a subtype of local order error, as seen in Table 5. In later versions of the assembly, the number of interleaving failures actually increased. Until now, all the algorithmic solutions have simply considered scaffolds to be non-overlapping, and have ignored this phenomenon. The order and orientation of the scaffolds themselves is correct, so these are considered separately from local errors.

We measured the base-pair accuracy of the sequence in the WGS assemblies. We found 46,722 discrepancies in WGS1, 26,355 in WGS2, and 13,095 in WGS3 corresponding to 3.99, 2.20, and 1.09 errors per 10,000 bp, respectively. The National Human Genome Research Institute (NHGRI) standard for finished sequence is less than one error per 10,000 bp. With the exception of gaps, WGS3 nearly meets this standard for finished sequence over the entire assembled sequence. However, most of the errors are in the reconstruction of repetitive sequence (Table 6); in the unique sequence, all three assemblies exceed the error rate standard.

Furthermore, assemblies tend to be less accurate at the tips of contigs, because the number of reads covering the ends is much lower than in the middle of the contig. As shown in Table 6, sequence quality rises when 10-50 bp are eliminated from the ends of each contig. Though the WGS sequences contain more gaps than would be considered acceptable for ‘finished’ sequence, the sequences that are present are highly accurate.

Conclusions

Generating a finished sequence

In producing Release 3, we have greatly improved the quality of the *Drosophila* euchromatic sequence. We have increased sequence accuracy and reduced the number of sequence contigs spanning the euchromatic portion of the genome from 1,100 to 50. The largest sequence contig covers the entire chromosome arm 3R in a single contig of 27 Mb. However, more remains to be done, and we have already begun work on Release 4. We are confident that all but a few of the remaining sequence gaps can be closed with existing technology. Likewise, we can easily improve the few regions of low sequence quality that remain. A greater challenge will be resolving complex tandem duplications, and discovering and correcting the sequence assembly errors that they cause. We have been fortunate to have access to two different assemblies, one of the whole genome, produced with a constrained assembler that utilizes paired-end information, and the other of BAC-sized intervals produced with an unconstrained assembler, phrap. Comparison of the products of these two assemblers with low-resolution restriction enzyme fingerprints of BAC clones has convinced us that constrained assemblies carried out on small genome intervals and more

Table 5

Order and orientation errors in the WGS assemblies compared to Release 3

	WGS1		WGS2		WGS3	
	Number of segments	Number of base-pairs	Number of segments	Number of base-pairs	Number of segments	Number of base-pairs
Aligned segments	2,125	113.30 Mb	2,270	114.41 Mb	1,087	114.99 Mb
Local errors*	9	68.33 kb	7	9.80 kb	3	5.64 kb
Interleaving failures†	17	39.42 kb	28	137.75 kb	33	139.42 kb
Repeat errors‡	25	42.52 kb	1	0.66 kb	1	0.98 kb
Gross misassemblies§	3	10.69 kb	0		0	

*Local errors include inversions and transpositions within a contig or that cause the order of contigs to be incorrect within a scaffold. †Interleaving failures are cases where it has not been recognized that two scaffolds overlap because the end contig in one scaffold lies in a gap in the adjacent scaffold. ‡Repeat errors are incorrect assemblies of transposable elements (see text for description). §Gross misassemblies are cases in which scaffolds themselves are out of order.

Table 6

Sequence error rates for the WGS assemblies

Errors per 10 kb	WGS1	WGS2	WGS3
All sequence	4.12	2.23	1.1
In tandem repeats	95.2	61.4	48.8
In interspersed repeats	78.2	15.8	9.62
In unique sequence	1.82	1.31	0.38
> 10 bp from gap	1.37	1.02	0.29
> 50 bp from gap	1.32	0.95	0.26

exact comparison to restriction digests are necessary to be truly confident of the fidelity of the sequence assembly, especially in repetitive regions.

WGS sequencing is efficient

In Release 2, we delivered a sequence of the euchromatic *Drosophila* genome that has proven extremely useful to experimental biologists, representing over 97.7% of the finished sequence. Repetitive sequences were of only draft quality, however. Using improved assembly software, 99% of the *Drosophila* genome was assembled accurately in WGS3, judged by comparison to the Release 3 sequence, with virtually no global errors and few local order and orientation errors. With the exception of a larger number of gaps, overall sequence quality approaches the NHGRI standard for finished sequence. Furthermore, it is likely that continued algorithmic advances and specific refinements to treat the case of tandem repeats will improve the quality of assembly in repetitive regions. Initial objections to the WGS strategy suggested that the finishing stage would be more difficult and expensive than BAC-by-BAC sequencing [25], in particular because of a lack of clones to use as finishing templates. However, we found that the 10-kb plasmids used in the WGS provide an excellent resource for finishing. Such clones are large enough to cover most gaps and small enough to sequence conveniently by transposon insertion methods.

If we were to start today?

In the course of the *Drosophila* Genome Project we have utilized a wide variety of sequencing strategies. On the basis of our experience, our strategy to sequence another genome would be the following. First, carry out WGS sequencing and assembly using paired-end reads from 2- and 10-kb plasmids and from 160-kb BAC clones at 7x, 5x and 0.17x sequence coverage, respectively. Second, align the BAC end sequences to the assembled sequence scaffolds to generate a preliminary physical map. Third, generate high-resolution restriction enzyme fingerprints of BAC clones in a three-deep tiling path across each arm to resolve collapsed repeats and to verify the physical map and final assembly. Fourth, localize a sampling of BACs using fluorescence *in situ* hybridization (FISH) to associate the sequence scaffolds to their chromosome of origin. Fifth, close gaps using the WGS 10-kb clones as sequencing templates. Note that we would not generate sequence information from the BACs, except for their end sequences, choosing instead to use the 10-kb WGS clones as templates for all finishing work. Sixth, improve low-quality regions with custom primers.

Materials and methods

Strain and libraries

Sequencing templates were made from P1, BAC and WGS DNA libraries using the *D. melanogaster* strain *yellow* (y^1); *cinnabar* (cn^1) *brown* (bw^1) *speck* (sp^1). This isogenic strain was constructed in the early 1990s [26]; the P1 [27], BAC [11] and WGS DNA libraries [1] were made in 1990, 1998 and 1999, respectively. Although we have not determined the single-nucleotide polymorphism rate between the libraries, we observed four cases of insertional polymorphisms in which BACs contain transposable elements that are absent from the Release 2 WGS assembly; two on the X, a *gtwin* element in BACR33A08 and a *412* element in BACR29P19; one on chromosome 2, a *roo* in BACR01K07 and one on chromosome 3, a *roo* in BACR02C22. We have confirmed the molecular mutation of the y^1 allele as an A to C transversion in the ATG translation initiation codon as first determined by Geyer *et al.* [28]. We determined the molecular lesion of the cn^1 allele to be a 1,832 bp deletion relative to wild type. The mutation in bw^1 was known to be associated with an uncharacterized insertion [29]. We have identified the insertion to be a *412* transposable element mapping to the third exon. The wild-type *sp* gene, located genetically and cytologically to 60C, has not yet been molecularly characterized. However, sp^1 is known to be suppressible by *suppressor of sable* [30], a known suppressor of *412* elements. Two *412* elements map to 60C, one in *Dat* and another near *Nop60B*.

Sequencing methods

BAC-based assembly

BAC-based assemblies were produced using phred version 0.000925.c [31] and phrap version 0.990329 [10]. WGS traces were obtained from Celera Genomics with a listing for each trace that included name, insert size and coordinates in Release 2. We determined the Release 2 coordinates of each BAC-end in our tiling path by comparison of the BAC-end sequences to the Release 2 sequence. The WGS sequence traces corresponding to the region spanned by each BAC in the tiling path were then pooled with the sequence traces generated from that BAC (approximately 1.5x coverage [1]) and assembled using phrap. At LBNL, draft sequence from neighboring BACs was also included in each assembly. Traces for repeat sequences, including transposable elements, were obtained from Celera Genomics in two sets: one (surrogate traces) with multiple location coordinates in Release 2, and another (800k traces) with no location coordinates. Only those sequence traces derived from repetitive

DNA that could be associated with their unique mate pairs were included in our assemblies. A total of 950 BAC clone assemblies were generated to complete the euchromatic BAC tiling path.

Sequence gap closure

The phrap assemblies were viewed using consed version 10 [32] and more recently, version 12.0. At LBNL, BAC end sequences were manually tagged, and if the assembly was in more than one contig, gap sizes were estimated using Consed and phrapview version 0.960731 [10]. To aid in the determination of gap sizes, phrap-based assemblies were compared to the Release 2 assembly using Sim4 [33]. At HGSC, gap sizes and locations were estimated, before phrap assembly, using the WGS read coordinates. Gap closure status was monitored in the phrap assemblies using check-contig.pl, a program developed at HGSC. Assemblies were automatically tagged using BLAST [34] or cross-match [10] to identify the location of transposable elements.

LBNL and the HGSC used slightly different strategies for gap filling. With an estimated gap size, LBNL used one of five gap-filling methods to determine their sequence: first, gaps less than 500 bp were sequenced using custom primers designed by consed and a 3-kb or 10-kb plasmid as DNA template; second, gaps between 500 bp and 2.5 kb that were spanned by a 3-kb clone were completed by sequencing the 3-kb clone using a transposon-based sequencing strategy, described below; third, gaps between 2.5 kb and 10 kb that were spanned by a 10-kb clone were completed by sequencing the 10-kb clone using a transposon-mediated sequencing strategy; fourth, gaps larger than 10 kb were completed by sequentially sequencing multiple 10-kb clones to walk across the gap. At LBNL, very rarely, a PCR product was generated and sequenced using custom designed primers.

HGSC's initial approach to gap closure of 3L was to design primer pairs to span each gap. One thousand base-pairs of sequence flanking each gap (2,000 bp total per gap) was extracted from the published chromosome sequence. The flanking sequences were masked for repeats using Repeatmasker [35] and then 16-22 bp PCR primer pairs were chosen for each gap using PRIMER 3 [36], such that the predicted melting temperature was 58 ± 3°C for gaps of less than 2,000 bp and 62 ± 6°C for gaps of more than 2,000 bp. We successfully generated PCR products for 290 gaps up to 9,000 bp. Internal sequencing primers were synthesized for sequencing PCR products less than 2,000 bp. PCR products greater than 2,000 bp were treated one of two ways. If the sequence of the product corresponded to a known transposon, a battery of custom-made sequencing primers evenly spaced across the element was used to generate the sequence. Alternatively, random shotgun libraries were generated and sequenced to 6x coverage. HGSC used phrap to assemble individual BACs and derive quality scores. Four BACs containing large, nearly exact, duplications refractory to assembly by phrap were resolved using the Euler assembler [37].

Sequencing 3- and 10-kb clones

LBNL and HGSC sequenced 10-kb clones using different strategies as part of the finishing process. LBNL employed a γδ transposon-based strategy to sequence 3-kb clones essentially as described [38]. An *in vitro* transposition system (Finnzymes TGS) was used to sequence 10-kb clones. Either 24 or 48 colonies per clone were selected for sequencing of 3 or 10-kb clones, respectively.

10-kb clones at HGSC were usually sequenced using shotgun libraries. Libraries were prepared according to the double adaptor method [39] with the following modifications: DNA was sheared using the Hydroshear (Gene Machines, San Carlos, CA) at speed code 2 for 25 cycles. The phenol extraction and ethanol precipitation after end repair and ligation steps were replaced with purification using Qiagen PCR Cleanup columns. Shotgun libraries of PCR products were constructed as the other libraries but without the size-selection step and with an increase in the shearing pressure from 10 to 20 psi and increase in time from 2 to 5 min. Some clones at HGSC were sequenced using the 'EZ::TN <KAN-2> Tn*5* transposon-based strategy [40].

Quality

BAC assemblies in a single sequence contigs were evaluated for base-pair accuracy using the phrap consensus quality scores. Regions of low quality were identified and sequenced using custom primers designed by Consed Autofinish [41] to bring the estimated error rate to less than one error in 30,000 bp. Completed high-quality BAC sequence was submitted to GenBank. Unfinished BACs were submitted to GenBank as phase 1 sequence. Tandem duplications were resolved by adding phd files from subassemblies. In cases of multiple identical copies of repeated sequence, phd files were used to generate an accurate assembly, but not necessarily an accurate quality score, as it is impossible to automatically constrain the location of individual traces. At LBNL, 340 BACs (40.9 Mb non-redundant sequence) had no difficult repeats and assembled easily using phrap. For the 101 BACs (13.6 Mb of non-redundant sequence) that included a finished sequence of a P1 clone, the assembly was driven by the P1 sequence. In order for phrap to correctly assemble 227 BACs (25.8 Mb non-redundant sequence) it was necessary to include phd files corresponding to either the sequence of a 3-kb (totaling 1.65Mb) or a 10-kb (totaling 2.3 Mb) plasmid, or other phd files (totaling 3.4 Mb) to obtain a correct assembly (see Repeat assembly below for more details).

Assembly verification

At LBNL, BAC assemblies were viewed using phrapview, which was customized to recognize our 2-kb, 3-kb and 10-kb subclones and display the paired-end relationships with a color-coded key. Clones that were beyond acceptable size or orientation parameters are visualized in red as 'problem clones'. Clusters of such clones indicate a collapsed repeat or misassembly. At HGSC, in addition to phrapview, BAC assemblies were verified by comparing the order of the reads in the Release 2 assembly with the order of the reads in the contigs generated using phrap and correlating the expected gap size with the completed sequence. This enabled rapid visualization of large-scale discrepancies, such as collapses or inversions, between the WGS and the phrap assemblies. Final BAC assemblies were then verified by comparing the virtual restriction digest to *Eco*RI, *Bam*HI and *Hind*III fingerprints produced by HGSC and manually called at LBNL and HGSC. Band sizes in the range of 1.6 to 16 kb were used for comparison. At LBNL, Release 3 BAC sequence assemblies were also compared to the Release 2 sequence from that region using Sim4 to analyze the large-scale assembly, and Sim3 to identify base-pair differences.

Repeat assembly

At LBNL we developed two custom assembly programs, assembleSubclone.pl and deconstruct.pl to facilitate resolution of tandem repeats. AssembleSubclone.pl was used to join the two traces, generated by primers in opposite direction from the same transposon insertion, prior to phrap assembly of 3- or 10-kb plasmid clones; this process effectively doubles the read length. A phd file generated from the assembly of the plasmid clone was used to direct the BAC assembly. Deconstruct.pl was used to resolve collapsed repeats provided that sufficient variation exists in the repeat copies. Using deconstruct.pl, we extracted local regions from the main assembly and reassembled the associated traces and their mate pairs at high stringency using phrap. The resulting consensus sequence was then incorporated in a reassembly of the entire BAC.

Arm assembly and overlap verification

Quality scores and consensus sequence from the BAC assemblies were exported from Consed. Chromosome arm assemblies were generated by determining the overlap between neighboring BACs using pslayout (LBNL) a program developed for this purpose as part of the human genome project [42] or by BLAST (HGSC). Discrepancies in overlap regions were reviewed and resolved. The programs Assemble-arms (LBNL) or BACstitcher (HGSC) was used to generate a FASTA file of the chromosome arms, starting with the first BAC in the tiling path and adding the unique sequence from the adjoining BAC to extend the sequence contig. To assess the accuracy of the chromosome arm assemblies we determined the location and orientation of the BAC-end sequences using a customized version of Sim4 [43]; we also checked that each gene present in the annotated Release 2 sequence was appropriately located in Release 3 using BLAST. To generate quality scores for an entire chromosome arm, we associated a phrap consensus score with each base. In regions of overlap between neighboring BACs, the highest consensus score was used. Gaps represented by Ns were not given a consensus quality score and did not contribute negatively to the estimated error rate. Sequence and physical map gaps are representing in the arm assembly by sets of Ns. One thousand Ns correspond to physical map gaps. If the gap size could be estimated, it was filled with the corresponding number of Ns.

At HGSC, unfinished BAC sequences were used to produce chromosome X (1-11) and chromosome 3L arm assemblies if the unfinished portion of the BAC was limited to one or two short (1-10 kb) regions of the clone but was high quality everywhere else. At HGSC, regions of the Release 2 sequence were used to produce chromosome X (1-11) and chromosome 3L arm assemblies if BAC clones spanning these regions had not yet been identified. In regions of overlap between finished clone and either unfinished or Release 2 sequence, BACstitcher preferentially incorporated finished sequence.

Sequence comparisons of Release 2 to Release 3

In order to quantify the base-pair differences between Release 2 and Release 3, we aligned the sequences using MUMmer v 2.1 [44]. We searched for maximal exact matches of at least 100 bp. From the list of exact matches, we generated a path of ordered and oriented matches that served as landmarks for subsequent analysis. In regions where discrepancies were detected, we compared Release 2 and Release 3 sequences using Sim4 [33] to determine subintervals of high similarity within the sequence and to count the number of individual base-pair differences within each subinterval.

Each unmatched segment was classified according to the nature of the mismatch as follows: known gaps in Release 2 which are filled with sequence in Release 3; newly discovered gaps in Release 2 defined as 20 bp or more of sequence that does not align with Release 3 using Sim4; and regions containing single nucleotide differences between Release 2 and Release 3, defined as less than 20 bp of sequence that does not align with Release 3 using Sim4.

Release 3 sequence was used to characterize the content of the known and newly discovered gaps in Release 2. The sequence content was annotated as: transposable element; homopolymer; simple repeats; tandem repeats; misassemblies; unique; gross misassemblies; and unfilled for those gaps that remain in Release 3. Transposable elements were identified using a curated list described in [9]. Homopolymers and simple repeats were determined using RepeatMasker. Tandem repeats were identified using Sim4 and the Release 3 gap sequence and the adjacent neighboring sequences. If the gap filling sequence showed 90% Sim4 similarity to the neighboring sequences, it was classified as a tandem repeat. Release 2 sequence that does not align with Release 3 was compared using Sim4 and the adjacent neighboring sequence. If the unaligned sequence showed 90%

sim4 similarity to the neighboring sequences, it was classified as a misassembly. Any interval not falling into one of these categories was considered unique. The seven gross misassemblies are described in Results and discussion.

Modifications to the WGS assembler

Between each assembly, the assembler was revised and improved. The basic algorithmic strategy for WGS is to first assemble the unique segments of the genome, and then resolve repeats. The assembler that produced WGS1 and the precursor to Release 2 used algorithms and code that had weaknesses with repeat resolution. Between WGS1 and WGS2, we corrected an error in the first step of repeat assembly - placement of repeat traces that are paired with an end read that lies in unique sequence. This allowed us to remove the second step in the original assembler's repeat reconstruction strategy, in which repeat traces were added to the assembly even when they were not paired with a unique read. This step was the cause of most of the inaccurate repeat reconstructions in Release 2. In addition, the overlap step of the assembler no longer requires explicit knowledge of genome-specific repeats. Between WGS2 and WGS3, the primary improvement was the introduction of a sequence error-correction algorithm that allowed us to improve the sequence quality of every read by examining its relation to all the other reads. This allowed us to tighten the stringency of overlap, leading to improved assembly, especially of repetitive regions.

Sequence comparisons of Release 3 to WGS 1, 2, and 3

The three assemblies were compared to Release 3 by first finding the best possible mapping of those sequence segments of the assembly that represent portions of the finished sequence. This task is complicated by nearly identical repeats that appear to align a portion of an assembly to every copy of the repeat. We proceeded in three steps. First, every contig was compared against the sequence of Release 3 using a local variation of MUMmer to produce a collection of significant aligned segments. Second, all of the scaffolds for which there are competing matches and are either under 50 kb and do not have at least 98% of their sequence matching, or are over 50 kb and do not have at least 60% of their sequence matching, were removed from consideration. What remains are scaffolds that clearly contain sequence belong to the euchromatin, with the exception of small scaffolds that were almost certainly entirely repetitive. Third, the remaining matched segments were then reduced by a greedy algorithm [45] to a one-to-one, non-overlapping sequence of matches that was considered to be the relevant tiling of Release 3 by each assembly. The set of scaffolds having a segment match in the one-to one tiling was considered to be mapped to Release 3. The heaviest weight common subsequence (HCS) between the WGS and Release 3 orders was computed where the weight of each match was its length; this order was taken as the correct order within each WGS assembly. Matched segments in the assemblies that were not in the HCS were considered order and orientation errors. We focused on the maximal set of matched segments to Release 3 without regard to their orientation and position on the contigs and scaffolds of the assembly. Thus we will see any inconsistencies between our assembly and Release 3. The content of the gaps in the WGS assemblies were identified using a repeat database and a tandem repeat finder [46] run with the default settings.

Acknowledgements

We thank Kate Nixon, Clare Doyle and Song Yang for help in the early gap closure phase of generating Release 3 sequence and Charles Yu, Hannibal Guarin and Ali Moshrefi for BAC subclone library construction. We thank Pieter de Jong and BACPAC resources for generating the *Eco*RI and sheared *D. melanogaster* BAC libraries. We thank Ben Berman and George Hartzell for software to align BAC end sequences to the different sequence releases. We thank Nomi Harris for software to assemble BAC sequence into chromosome arms, and Chris Mungall for verifying the gene content of the Release 3 assemblies. We are indebted to David Gordon for his continual modifications to Consed. We thank Harald Biessmann for helpful discussions concerning telomeres. This work was supported by Celera Genomics, the Howard Hughes Medical Institute and NIH grant P50-HG00750 (GMR) and carried out under Department of Energy Contract DE-AC0376SF00098, University of California.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
2. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, *et al.*: **A whole-genome assembly of *Drosophila*.** *Science* 2000, **287:**2196-2203.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
4. Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, *et al.*: **A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome.** *Science* 2002, **296:**1661-1671.
5. Hoskins RA, Smith CD, Carlson J, Carvalho BA, Halpern A, Kennedy C, Kaminker JS, Mungall C, Sullivan BA, Sutton G, *et al.*: **Heterochromatic sequences in a *Drosophila* whole genome shotgun assembly.** *Genome Biol* 2002, **3:**research0085.1-0085.16.
6. Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, Harris NL, 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.
7. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell K, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, *et al.*: **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biol* 2002, **3:**research0083.1-0083.22.
8. Lewis SE, Searle SMJ, Harris NL, 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.
9. Kaminker JS, Bergman C, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DL, 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
10. **The Phred/Phrap/Consed system home page** [http://www.phrap.org]
11. 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.

12. Peter A, Schottler P, Werner M, Beinert N, Dowe G, Burkert P, Mourkioti F, Dentzer L, He Y, Deak P, *et al.*: **Mapping and identification of essential gene functions on the X chromosome of *Drosophila*.** *EMBO Rep* 2002, **3:**34-38.
13. Locke J, Podemski L, Aippersbach N, Kemp H, Hodgetts R: **A physical map of the polytenized region (101EF-102F) of chromosome 4 in *Drosophila melanogaster*.** *Genetics* 2000, **155:**1175-1183.
14. Marra M, Kucaba T, Dietrich N, Green E, Brownstein B, Wilson R, McDonald K, Hillier L, McPherson J, Waterston R: **High-throughput fingerprint analysis of large-insert clones.** *Genome Res* 1997, **7:**1072-1084.
15. **FlyBase GadFly genome annotation database** [http://www.fruitfly.org/cgi-bin/annot/query]
16. Lifton RP, Goldberg ML, Karp RW, Hogness DS: **The organization of the histone genes in *Drosophila melanogaster*: functional and evolutionary implications.** *Cold Spring Harb Symp Quant Biol* 1978, **42:**1047-1051.
17. Tycowski KT, Steitz JA: **Non-coding snoRNA host genes in *Drosophila*: expression strategies for modification guide snoRNAs.** *Eur J Cell Biol* 2001, **80:**119-125.
18. Karpen GH, Spradling AC: **Analysis of subtelomeric heterochromatin in the *Drosophila* minichromosome Dp1187 by single P element insertional mutagenesis.** *Genetics* 1992, **132:**737-753.
19. Gatti M, Pimpinelli S: **Functional elements in *Drosophila melanogaster* heterochromatin.** *Annu Rev Genet* 1992, **26:**239-275.
20. O'Hare K, Chadwick BP, Constantinou A, Davis AJ, Mitchelson A, Tudor M: **A 5.9-kb tandem repeat at the euchromatin-heterochromatin boundary of the X chromosome of *Drosophila melanogaster*.** *Mol Genet Genomics* 2002, **267:**647-655.
21. Carvalho BA, Vibranovski MD, Carlson J, Celniker SE, Hoskins RA, Rubin GM, Sutton G, Adams MA, Myers EW, Clark AG: **Y chromosome and other heterochromatic sequences of the *Drosophila melanogaster* genome; how far can we go?** *Genetica*, in press.
22. Eichler EE: **Recent duplication, domain accretion and the dynamic mutation of the human genome.** *Trends Genet* 2001, **17:**661-669.
23. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler.** *Genome Res* 2002, **12:**177-189.
24. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, *et al.*: **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*.** *Science* 2002, **297:**1301-1310.
25. Green P: **Against a whole-genome shotgun.** *Genome Res* 1997, **7:**410-417.
26. Brizuela BJ, Elfring L, Ballard J, Tamkun JW, Kennison JA: **Genetic analysis of the *brahma* gene of *Drosophila melanogaster* and polytene chromosome subdivisions 72AB.** *Genetics* 1994, **137:**803-813.
27. Smoller DA, Petrov D, Hartl DL: **Characterization of bacteriophage P1 library containing inserts of *Drosophila* DNA of 75-100 kilobase pairs.** *Chromosoma* 1991, **100:**487-494.
28. Geyer PK, Green MM, Corces VG: **Tissue-specific transcriptional enhancers may act *in trans* on the gene located in the homologous chromosome: the molecular basis of transvection in *Drosophila*.** *EMBO J* 1990, **9:**2247-2256.
29. Dreesen TD, Johnson DH, Henikoff S: **The brown protein of *Drosophila melanogaster* is similar to the white protein and to components of active transport complexes.** *Mol Cell Biol* 1988, **8:**5206-5215.
30. Searles LL, Voelker RA: **Molecular characterization of the *Drosophila vermilion* locus and its suppressible alleles.** *Proc Natl Acad Sci USA* 1986, **83:**404-408.
31. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8:**175-185.
32. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8:**195-202.
33. 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.
34. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, D.J. L: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
35. **RepeatMasker documentation** [http://ftp.genome.washington.edu/RM/RepeatMasker.html]
36. **Primer3 software distribution** [http://www-genome.wi.mit.edu/genome_software/other/primer3.html]
37. **Euler** [https://gridport.npaci.edu/euler]
38. Kimmel B, Palazzolo M, Martin CH, Boeke JD, Devine SE: **Transposon-mediated DNA Sequencing.** In *Genome Analysis*, Vol 1, Edited by Birren B, Green E, Klapholz S, Myers RM, Roskams J. New York: Cold Spring Harbor Laboratory Press; 1997:455-532.
39. Andersson B, Wentland MA, Ricafrente JY, Liu W, Gibbs RA: **A "double adaptor" method for improved shotgun library construction.** *Anal Biochem* 1996, **236:**107-113.
40. Goryshin IY, Reznikoff WS: **Tn5 *in vitro* transposition.** *J Biol Chem* 1998, **273:**7367-7374.
41. Gordon D, Desmarais C, Green P: **Automated finishing with autofinish.** *Genome Res* 2001, **11:**614-625.
42. Kent WJ, Haussler D: **Assembly of the working draft of the human genome with GigAssembler.** *Genome Res* 2001, **11:**1541-1548.
43. Hartzell GW: *An assessment of genome annotation tools and an approach to solving a set of problems from a genome sequencing project.* PhD thesis. Berkeley: University of California, Berkeley; 2001.
44. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic Acids Res* 2002, **30:**2478-2483.
45. Halpern A, Huson DH, Reinert K: **Segment match refinement and applications.** In *Algorithms in Bioinformatics, Proceedings of the Second International Workshop.* Edited by Guigo R, Gusfield, D. Heidelberg: Springer; 2002: 126-139.
46. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
47. Sorsa V: *Chromosome Maps of* Drosophila. Boca Raton, FL: CRC Press; 1988.
48. **BACPAC resources center** [http://www.chori.org/bacpac]
49. **G16 Finishing Standards for the Human Genome Project - Version September 7, 2001** [http://genome.wustl.edu/Overview/finrulesname.php?G16=1]
50. Madueno E, Papagiannakis G, Rimmington G, Saunders RD, Savakis C, Siden-Kiamos I, Skavdis G, Spanos L, Trenear J, Adam P, *et al.*: **A physical map of the X chromosome of *Drosophila melanogaster*: cosmid contigs and sequence tagged sites.** *Genetics* 1995, **139:**1631-1647.
51. Benos PV, Gatt MK, Murphy L, Harris D, Barrell B, Ferraz C, Vidal S, Brun C, Demaille J, Cadieu E, *et al.*: **From first base: the sequence of the tip of the X chromosome of *Drosophila melanogaster*, a comparison of two sequencing strategies.** *Genome Res* 2001, **11:**710-730.