REVIEW

# A Whole-Genome Assembly of *Drosophila*

Eugene W. Myers,[1]* Granger G. Sutton,[1] Art L. Delcher,[1] Ian M. Dew,[1] Dan P. Fasulo,[1] Michael J. Flanigan,[1] Saul A. Kravitz,[1] Clark M. Mobarry,[1] Knut H. J. Reinert,[1] Karin A. Remington,[1] Eric L. Anson,[1] Randall A. Bolanos,[1] Hui-Hsien Chou,[1] Catherine M. Jordan,[1] Aaron L. Halpern,[1] Stefano Lonardi,[1] Ellen M. Beasley,[1] Rhonda C. Brandon,[1] Lin Chen,[1] Patrick J. Dunn,[1] Zhongwu Lai,[1] Yong Liang,[1] Deborah R. Nusskern,[1] Ming Zhan,[1] Qing Zhang,[1] Xiangqun Zheng,[1] Gerald M. Rubin,[2] Mark D. Adams,[1] J. Craig Venter[1]

**We report on the quality of a whole-genome assembly of *Drosophila melanogaster* and the nature of the computer algorithms that accomplished it. Three independent external data sources essentially agree with and support the assembly's sequence and ordering of contigs across the euchromatic portion of the genome. In addition, there are isolated contigs that we believe represent nonrepetitive pockets within the heterochromatin of the centromeres. Comparison with a previously sequenced 2.9-megabase region indicates that sequencing accuracy within nonrepetitive segments is greater than 99.99% without manual curation. As such, this initial reconstruction of the *Drosophila* sequence should be of substantial value to the scientific community.**

The primary obstacle to determining the sequence of a very large genome is that, with current technology, one can directly determine the sequence of at most a thousand consecutive base pairs at a time. The process, dideoxy sequencing, used to produce such sequencing reads was essentially invented by Sanger circa 1980 (*1*), with subsequent modest gains in read length, moderate gains in data accuracy, and significant gains in throughput. Given the limitation on read length, researchers employ a shotgun-sequencing approach, in which an effectively random sampling of sequencing reads is collected from a larger target DNA sequence. With sufficient oversampling, the sequence of the target can be inferred by piecing the sequence reads together into an assembly.

Early on, the shotgun approach was applied to small viral genomes and to 30- to 40-kbp segments of larger genomes that could be manipulated and amplified in a cosmid. For a given level of oversampling, the number of unsampled regions or gaps increases linearly with target size, as does the number of interspersed repetitive sequences that tend to confound assembly. After computer assembly, a finishing phase ensues, wherein the gaps between assembled contigs are closed experimentally, and any misassembly is resolved. Because one does not know the order of the contigs or the size of the gaps and because the assembly problem becomes harder as size increases, it was commonly believed that cosmid targets represented the limit of the shotgun approach. Whole genomes were sequenced by first developing a set of cosmids or other clones covering the genomes by a process called physical mapping, and then shotgun sequencing each clone as in (*2*–*4*).

In 1994, the sequence of *Haemophilus influenzae* was obtained from the assembly of a whole-genome data set obtained by shotgun sequencing (*5*). This bacterial genome, at 1.8 Mbp, was much larger than was previously thought possible by a direct shotgun approach, the largest previous genome so sequenced being the lambda virus in 1982 (*6*). Critical to this accomplishment was the construction of a computer program capable of performing the assembly and the use of pairs of reads, called mates, from the ends of 2-kbp and 16-kbp inserts randomly sampled from the genome. Even though the pairing information was false 10 to 20% of the time owing to lane tracking problems on the slab-gel sequencing instruments available at the time, the presence of several mates with one read in one contig and the other read in another contig allowed ordering of the contigs and gave a rough estimate of the size of the gap between them, simplifying the finishing phase. Many groups have since sequenced bacterial genomes this way, and investigators have moved from using shotgun sequencing for cosmids to targets on the order of 100 to 150 kbp, that is, those clonable in P1 and bacterial artificial chromosome (BAC) vectors.

A new approach to sequencing whole genomes, proposed by Venter, Smith, and Hood in 1996 (*7*), was to make a 15× library of BAC-sized inserts randomly sampled from the genome and to produce end-sequence read pairs for them. One could then select and apply shotgun sequencing to several seed BACs, whereupon the end-sequences of other BACs in the library could be used to determine minimally overlapping BACs at each end to sequence next in an interactive walk across the genome. Weber and Myers then proposed the whole-genome shotgun sequencing of the human genome in 1997 (*8, 9*). The protocol involved collecting a 10× oversampling of the genome, with mate pairs from 0.9-kbp and 10-kbp inserts in a 1:1 ratio, and assembling this in conjunction with the long-range information provided by a genome-wide sequence-tagged site (STS) map that is a series of unique, 300- to 500-bp sites ordered across the genome with an average spacing between sites of 100 kbp. In 1998, Venter and colleagues announced the undertaking of a whole-genome shotgun sequencing of the human genome (*10*) with the sequencing of *Drosophila* serving as a pilot project.

For *Drosophila*, we set about collecting a 10× oversampling of a genome using a 1-to-1 ratio of 2-kbp and 10-kbp mate pairs. In addition, enough BACs to provide 15× coverage of the genome were to be collected and end-sequenced, effectively generating a set of mate pairs that give long-range information similar to that provided by the STS maps described above. *Drosophila*'s euchromatic genome is estimated at 120 Mbp. Thus, the protocol would require collecting at least 2.4 million reads and 15,000 BAC end pairs, totaling 1.2 billion base pairs of data. Our *Drosophila* sequencing project began in May 1999, in collaboration with the Berkeley *Drosophila* Genome Project (BDGP), the results of which are detailed in the accompanying papers (*11, 12*).

## Celera Assembler Design Principles

The primary difficulty in building an assembler for a whole-genome shotgun data set is to develop an algorithmic approach that detects and is not confused by stretches of repetitive DNA. The key to not being confused by repeats is the exploitation of mate pair information to circumnavigate and to fill them (*13*). Because of this, the result of assembly is a set of scaffolds of contigs, versus a set of contigs as customarily produced by other assemblers. A scaffold is a set of contigs that are ordered, oriented, and positioned with respect to each other by mate pairs whose reads are in adjacent contigs (see below). Although we demonstrate below that a whole-genome shotgun data set can be assembled in isolation, our pragmatic objective

[1]Celera Genomics, Inc., 45 West Gude Drive, Rockville, MD 20850, USA. [2]Howard Hughes Medical Institute, Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be sent: Gene.Myers@celera.com

is to produce the best possible reconstruction of a genome, along with its correlation to existing data. Therefore, the assembler is capable of utilizing available external data. Our assembler places reads in a series of stages, starting with the safest "moves" and progressing toward increasingly more aggressive ones. The stage and evidence for a read's placement are open to inspection, providing an audit trail of the assembler's decision-making. To further optimize development time, we decided to build a batch assembler that assumes all data are available when it begins its task. For *Drosophila* this was feasible because assembly of a complete data set takes less than a week on an eight-processor suite of Compaq Alpha ES40s with a 32-Gb memory (*14*).

## The *Drosophila* Data Sets

The scale of whole-genome assembly dictates that the quality of the input data be much higher than that required for smaller assembly problems. We determined data requirements on the basis of simulation estimates (*15*) and received data of the quality shown in Table 1. In a whole-genome context, trillions of overlaps between reads are examined. In order to keep the a posteriori probability of a false overlap low, regions of low sequence quality must be trimmed much more aggressively than for other protocols (*16*). We produced 3.156 million reads that yielded 1.76 Gbp of sequence after trimming to the 98% accuracy level on the basis of quality values that reflect the log-odds score of the base's being correct (*17*). The observed mean sequencing accuracy of these reads after trimming was 99.5% (*18*).

A substantial fraction of the reads must be in mate pairs if one expects to achieve long-range ordering and repeat filling. Moreover, the more accurately one knows the distance between a pair of mates, and the more reliably one knows that a given pairing is true, the more strongly one can make inferences during the assembly process. We produced 1.151 million pairs (72.8% of the reads) whose insert lengths were normally distributed with 10% variance and whose pairing reliability has been estimated at 99.66% (*19*).

The spectrum of 2-kbp, 10-kbp, and BAC mates must be such that all of the euchromatic, nonrepetitive DNA is linked together and covered at least two deep at every point. Moreover, an insufficient number of 10-kbp and BAC mates will prevent the formation of assemblies covering each chromosome arm. To our surprise, 10-kbp inserts could be sequenced as successfully as 2-kbp inserts, so we increased production of the 10-kbp mates in the late stages to produce 654,000 of the 2-kbp mates and 497,000 of the 10-kbp mates. A total of 12,152 acceptable quality BAC mates of average separation 130.2 kbp, generated at Genoscope (*11*), were received from the BDGP and European *Drosophila* Genome Projects (EDGP).

We term the data set described above the whole-genome shotgun data set or WGS data set, as it provides the data stipulated in our pure conception of the whole-genome shotgun sequencing protocol. In addition to these data, the BDGP constructed a map of the second and third chromosomes, completely sequenced 340 BAC and P1 inserts comprising about 26 Mbp of *Drosophila* euchromatic sequence, and produced a 1.28× draft shotgun of each BAC and P1 clone in a tiling set chosen from a physical map covering the genome (*20*). The EDGP produced a map of the X chromosome and completely sequenced cosmid and BAC clones covering about 3 Mbp. The Canadian *Drosophila* Genome Project produced a physical map of the small fourth chromosome. For more details on these data sets, see table 1 of (*11*). The joint data set is our term for the WGS data plus the draft reads and a perfect shredding (*21*) of 340 of the completely sequenced clones into a 3× tiling of 550-bp reads. There were a total of 337,000 draft reads constituting 153.1 Mbp of sequence and 154,000 reads shredded from the completed BACs. We did not include the known STS markers for *Drosophila* in the joint data set, reserving them for independent confirmation, and no specific advantage was taken of the locality and ordering of the included external data. Thus, the net effect is that each of these reads was simply mapped to a location in the assembly, possibly filling in sequence gaps by means of the extra sampling coverage they provided. The primary use of these marker reads was to validate assembly and to provide navigation information for the finishing stage.

The finished sequence resulting from assembly of the joint data set along with current finishing efforts will be available both at Celera's Web site and at GenBank under accession numbers AE002566–AE003403 An assembly of the data through the scaffolding phase (see below) was deposited in GenBank on 31 December 1999, accession numbers AC012691–AC020545. We are also prepared to participate in appropriate collaborative efforts to use our raw data to test future algorithms.

## Celera Assembler's Algorithmic Design

The Celera assembler consists of a pipeline of several stages as shown in Fig. 1. An illustrated primer on the assembler algorithms is on the Web (www.celera.com/genomeassembler). In preparation for the assembly computation, the electropherograms for a read were interpreted as a sequence of bases and associated quality values that reflect the log-odds score of the base's being correct (*17*). The read was then trimmed to an interval of 98% accuracy according to these quality values. Any prefix sequence of the high-quality region matching the sequencing vector or linker was aggressively removed. Finally, the remaining portion of the reads were screened for matches to any contaminant DNA such as *Escherichia coli* or cloning or sequencing vectors, and the entire read was removed if a significant matching segment was found. After this processing, what remained was a set of high-quality reads of *Drosophila* se-

**Table 1.** Input data requirements and characteristics. The requested column gives the minimum or maximum requirement for the item stipulated at the left of each row. The received column shows what was actually produced.

| Type of data | Requested | Received |
|---|---|---|
| Read length and accuracy | 500 bp @ 98% (min) | 551 bp @ 98% |
| Shotgun coverage | 10× (min) | 14.6× |
| Reads in pairs | 70% (min) | 72.8% |
| Insert length variance | ±3% (max) | ±10% |
| False-positive pairs | 1% (max) | 0.34% |
| BAC map coverage | 15× (min) | 13.18× |
| Ratio of 2 kbp to 10 kbp | 4 to 1 (max) | 1.32 to 1 |

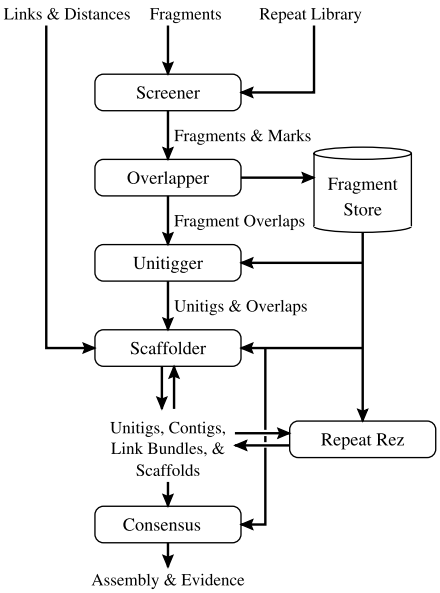


**Fig. 1.** Assembly pipeline. From an engineering perspective, sequences of messages flow from one stage to the next. Each stage performs work on its input stream, producing a stream of output messages reflecting its transformational function. The text gives the function of each stage.

quence, namely, fragments.

*Screener*. Each input fragment was checked for matches to known repetitive elements, either noting matched regions, a soft screen, or masking them from further consideration, a hard screen. For *Drosophila*, the library of known repetitive elements was a manually curated list of its ribosomal DNA, histones, heterochromatin, and known retrotransposons. We chose to hard screen matches to ribosomal and heterochromatic DNA. This implies that these portions of the genome would not be assembled, because overlaps interior to masked regions were not computed. However, this is consistent with the implicit goal of all sequencing efforts, that is, to determine the sequence of the euchromatic segments of the genome.

Only 2.50% of the sequence matched heterochromatin, and almost all such matches covered only part of a read, confirming that heterochromatic sequence does not clone in our larger inserts. For the other hard-screened items, 3.01% of the sequence matched ribosomal sequence, 0.38% matched histones, and 0.13% was microsatellite sequence found by a de novo low-complexity sequence detector. Retrotransposons matched 7.26% of the incoming sequence, and 1.48% matched other known moderate repeats. Unfortunately, we had to hard screen 1.51% of the data matching a retrotransposon found in the ribosomal DNA, and we conjecture that this may be the cause of several repeat-sized gaps remaining in our assembly. In total, 7.53% of the data were hard screened and 8.74% were soft screened.

*Overlapper*. Each fragment was compared with all fragments previously examined in search of overlaps with fewer than 6% differences and involving at least 40 bp of unmasked sequence. Any overlap meeting this criterion must be either true or repeat-induced (Fig. 2). Our methodology is similar to the seed-and-extend idea developed for BLAST (*22*), save that our implementation, tuned for high-stringency matches, compares 32 million pairs of reads every second. Even so, the total CPU time required mandated the use of parallel processing. The overlapper was organized to compare two batches of sequences, taking care not to compare reads against themselves if the two batches happened to be the same set of sequences. With this simple distributed architecture and a controlling program to collate results, the computation could be spread across as many processors as desired.

For the WGS data set, 212 million overlaps were computed for an average of 33.7 overlaps per fragment end. However, this is misleading, as one has essentially a Poisson distribution with mean 13.7 and a very long tail of fragments with up to 4000 overlaps at a given end. The fragments with very large numbers of overlaps are clearly portions of repeats.

*Unitigger*. Collections of fragments whose arrangement is uncontested by overlaps from other fragments were assembled into what we call unitigs. Each unitig was assessed as to whether it represented unique or repetitive sequence. Those certain to represent unique DNA were designated U-unitigs. Potential boundaries of repeat sequences were sought at the tips of the U-unitigs, and those found were used to extend U-unitig ends as far as possible into a repeat.

Mathematically, a unitig is a maximal interval subgraph of the graph of all fragment overlaps for which there are no conflicting overlaps to an interior vertex. This idea was originally explored by Myers (*23*) and extended by us to treat the case in which one read entirely matches a subsegment of another (Fig. 3). After this step, one goes from 3.158 million fragments to 54,000 unitigs with two or more fragments, and from 221 million overlaps to 3.104 million between unitigs: 48- and 68-fold reductions in problem size, respectively.

Almost every unitig is a correct subassembly of fragments. The exception occurs when a set of reads sampled from the interior of copies of a very high fidelity repeat ($X' + X''$ in Fig. 3) overcollapses into a unitig because they all form a consistent subassembly of the repeat's interior. We detect these unitigs by computing an A-statistic that is the log-odds ratio of the probability that the distribution of fragment start points is representative of a correct versus an overcollapsed unitig of two repeat copies (*24*). In all of our simulation runs, including synthetic genomes as large as 100 Mbp, we never encountered an incorrectly assembled unitig with a score greater than 10. We term unitigs with an A-statistic greater than 10 "U-unitigs" as they almost certainly represent unique DNA in the genome that has been correctly assembled. We found 9413 U-unitigs with an average length of 12.2 kbp and totaling 115.4 Mbp of sequence.

By detecting repeat boundaries, we could identify and remove some of the repetitive overlaps between unitigs. Whenever a unitig *A* overlaps two unitigs *B* and *C* at one end, then by construction the initial portions of *B* and *C* align, but at some point *B* and *C* fail to overlap and we can find this repeat boundary accurately with dynamic programming. We found 8570 repeat boundaries in the WGS data set and simulations support the conclusion that they represent 90% of all such boundaries. Any overlap from U-unitig *X* to unitig *Y* entirely on the repeat side of a boundary can safely be eliminated if there is another overlap, not so contained, whose destination does not overlap *Y*. This enables further U-unitig extension, on the order of a read length, into a repetitive region. Repetitive elements shorter than the average read length were effectively resolved. After this process, the number of U-unitigs reduces to 8389, and their average size increases by 1.7 kbp to 13.9 kbp, for a total of 116.3 Mbp in U-unitigs.

*Scaffolder*. All possible U-unitigs with mutually confirming pairs of mates or BAC ends were linked into scaffolds consisting of a set of ordered, oriented contigs for which the size of the intervening gaps is approximately known (Fig. 4). When the left and right reads of a mate are in different unitigs, their distance relation orients the two unitigs and provides an estimate of the distance be-

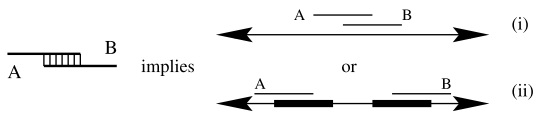


**Fig. 2.** True and repeat overlaps. Consider two fragments *A* and *B* that overlap as shown at left. There are two possible conclusions depicted at right: (i) the fragments were sampled from overlapping segments of the genome and so belong together in an assembly, a true overlap, or (ii) the overlapping portion is part of a repeated sequence that occurs multiple times in the genome, and the two reads do not belong together, a repeat overlap. Assembly would be a trivial matter if we could divine all the true overlaps; the key objective is to conservatively find true overlaps and to avoid the repetitive ones, especially early in the assembly process.

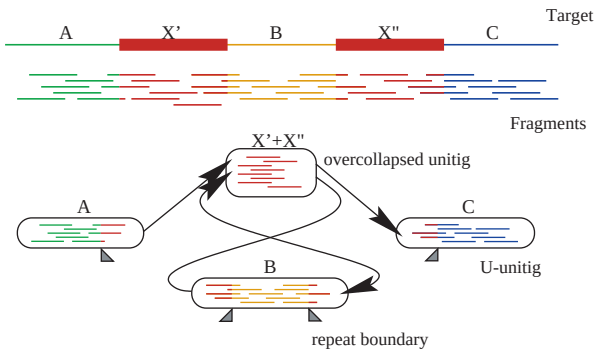


**Fig. 3.** Unitigs and repeat boundaries. Consider the hypothetical genome consisting of three unique stretches *A*, *B*, and *C* with two nearly identical, interspersed copies, $X'$ and $X''$, of a repeat element *X*. This results in the four unitigs and overlaps shown. As explained in the text the unitig $X' + X''$ is overcollapsed, and the U-unitigs for regions *A*, *B*, and *C* have repeat boundaries indicating the tail portions that project into *X*.

tween them. Unfortunately, this relation is false 0.34% of the time, and so one cannot trust the given inference. However, if two or more mates consistently indicate a given orientation and separation between two U-unitigs, the inference is estimated to be wrong only 1 in $10^{15}$ times. We first found bundles of mate pairs and overlaps that consistently place unitigs relative to each other. When these bundles had several contributing links, we computed a tighter expected average distance and deviation between the unitigs, especially when an overlap between them was part of the bundle. There were approximately 20,000 confirmed bundles between unitigs averaging 10.6 mate pairs per bundle.

In analogy to the unitigger, all sets of U-unitigs that were consistently ordered and placed by confirmed bundles, that is, containing two or more 2-kbp or 10-kbp links, were assembled into a scaffold of contigs where a contig is, at this stage, a series of overlapping U-unitigs. We then ordered and placed these scaffolds using a best-first selection of BAC bundles (that is, one involving a BAC mate) ordered on the number of links in the bundle. The normal distribution distance estimates between contigs were then refined on the basis of a least squares estimation by using all link estimations consistent with the scaffolding. The 24,000 bundles among the 8391 U-unitigs were distilled by the scaffolding into 3736 contigs of average size 30,631 bp with 5973 bundles between contigs supporting their order. At the end of this step we essentially had the euchromatic, nonrepetitive portion of the genome assembled and ordered.

*Repeat resolution: rocks, stones, pebbles*. Both intra- and interscaffold gaps were filled in a series of three, increasingly more aggressive, levels of repeat resolution. The rock-phase placed unitigs that were consistently positioned by at least two mate pairs, the stone-phase placed unitigs that were positioned by a single mate pair and confirmable by an overlap tiling across the gap containing it, and the pebble-phase attempted to find the best tiling across gaps using a quality-value based measure of significance.

Rocks are unitigs that have a positive A-statistic and have either two mate links that consistently link it to contigs on one or both sides or four or more links, where at most one does not agree with the others. In simulations, rock placements were always correct. For WGS data, 2827 rocks of average length 1035 bp were placed, closing 667 gaps of average width 457 bp and providing 1.70 Mbp of new assembled sequence.

All remaining unitigs have no confirmed bundles linking them to the assembly scaffold. Stones have only a single mate link to a contig on one side or another of a gap, but we further require that there be an overlap-based tiling of unitigs that fills the gap and includes the stone. The tiling path supports the stone, and we found such placement to be erroneous rarely in simulations, and only when the stone was so close to the sequence of the repeat copy that the impact on the accuracy of the reconstructed sequence was minimal. For WGS data, 160 stones of average length 1611 bp were placed, closing 77 gaps of average width 1327 bp and providing 144 kbp of new assembled sequence.

We then proceeded to find the best overlap tiling of unitigs across each gap, where any existed. As our measure of goodness, we used a log-odds ratio of the probability that an overlap is true versus repeat-induced on the basis of the quality values for the sequences. Some fragments were misplaced at this point, either because of following the incorrect path or using undetected overcollapsed unitigs. This occurred usually when a repeat was long, such as the full-length, 7- to 9-kbp retrotransposons of *Drosophila*, and its interior had to be constructed entirely from a pebble tiling. In general, however, the quality of these interior repeat segments was still better than 99.5% accurate. The discussion below comparing repeats in the Adh region further illustrates the nature of the errors incurred with long-repeat interiors. For WGS data, 30,998 pebbles of average length 640 bp were placed, closing 1257 gaps of average width 2219 bp and providing 3.21 Mbp of new assembled sequence. At this point, contigs average 50,002 bp in size.

*Consensus*. Reads were multiply aligned according to the consensus metric and consensus base calls were derived in the alignment columns. The quality of each consensus base was computed as the log-odds of correctness by using the quality values available for each read base.

The quality of the trimmed sequence in Celera's data is so high that a simple shift-and-evaluate algorithm we call "abacus" suffices to compute the optimal consensus-measure sequence. We then evaluated each column using a Bayesian estimate as described in earlier work (*25*). In particular, the consensus estimator will report positions that appear polymorphic with an estimate of the likelihood of the polymorphism being real, as opposed to error-induced.

Our assembler only uses quality values to drive the final pebble walks and to provide consensus quality values. All other decisions are made with percent sequence identity as the discriminating measure. This is a significant departure from the prevailing paradigm for assemblers (*26*).

## Characteristics of the *Drosophila* Assembly

The assembly of the joint data set resulted in 838 firm scaffolds, where we define a scaffold as firm if it contains at least one U-unitig. By definition all scaffolds that are not firm are unitigs with an A-statistic less than 10, and almost without exception, these unitigs are (i) unrelated to the firm scaffolds by either link or overlap relations, (ii) localized to repeat-induced gaps in the firm scaffolds, or (iii) pebbles that were relevant but not used in late-stage repeat resolution. We thus consider these firm scaffolds to be the result of assembly. For the firm scaffolds, 50 could be mapped to the euchromatic genome via markers of the BDGP STS map, and 134 could be mapped to the euchromatic genome via the draft sequencing of the BDGP physical map. In Adams *et al.* (*11*), these 134 mapped scaffolds are considered the preliminary reconstruction of the euchromatic sequence, wherein there are 1630 gaps to be finished. The remaining 704 scaffolds are all comparatively small U-unitigs with no observable connection to the draft or STS data, and we conjecture that some substantial fraction of these must be nonrepetitive islands within the heterochromatin of the centromeres (*11*), or may represent as-yet-unidentified foreign DNA.

There are a total of 119.71 Mbp of sequence in the 2483 contigs of the firm scaffolds that span 122.76 Mbp when one allows for the estimated amount of sequence that is

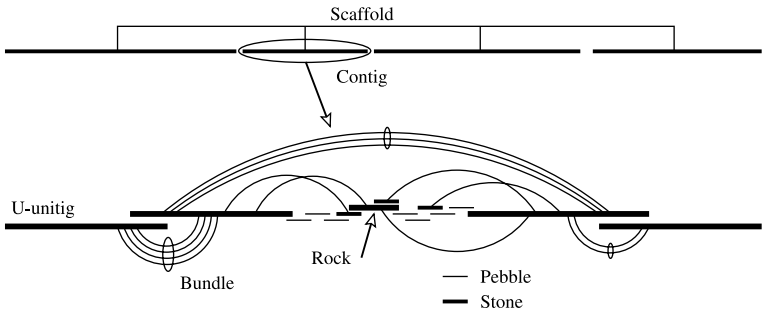


**Fig. 4.** Anatomy of a scaffold. A scaffold is a collection of ordered contigs with approximately known distances between them. Our contigs are built from U-unitigs that form a scaffold via bundles and then have a series of rocks, stones, and pebbles filled into the gaps between them (where possible).

in the gaps between a scaffold's contigs. Only 0.34% of the mated reads within contigs did not agree with the placement of their mates, which is well within the expected false-positive error rate of the pairing information. There are 70 scaffolds having spans over 30 kbp, and the 25 scaffolds with spans larger than 100 kbp contain more than 95% of the assembled sequence (114.1 Mbp). The sizes, in millions of base pairs, of the scaffolds over 1 Mbp are 24.3, 16.4, 15.1, 13.7, 10.6, 9.1, 5.1, 4.8, 4.5, 2.7, 2.1, 1.4, 1.4, and 1.3. These megascaffolds are a subset of the 50 mapped to the euchromatin by STS markers and cover the preponderance of the euchromatic portions of every chromosome arm, breaking up into smaller scaffolds as the telomeres and centromeres are approached. This can be seen in the segmentation of Fig. 5 where each segment represents a scaffold. It was simplest for us to arrange our investigation around the size of a scaffold, so in the remainder of this section we discuss the nature of these 25 scaffolds. The qualitative features to be observed about these scaffolds are representative of the entire set.

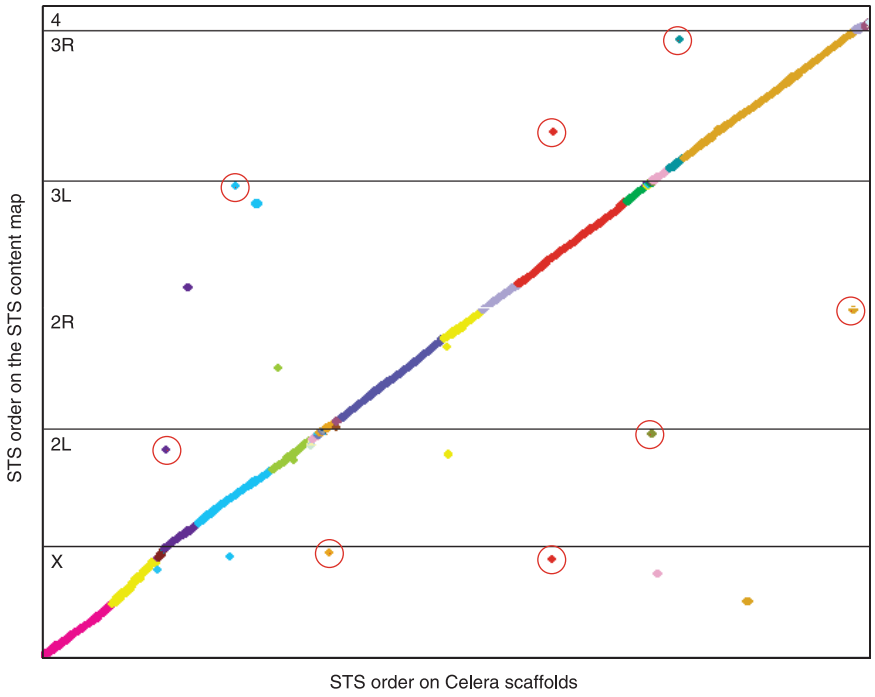


**Fig. 5.** STS-content map. Celera assembly scaffolds were plotted against the STS order on the STS-content map. The color palette is used to delineate scaffolds. The 17 outliers have been investigated; those circled in red remain unresolved at the time of publication.

**Table 2.** Comparison of assembly of scaffolds larger than 100 kbp on three data sets. The length column for the scaffolds row gives the total number of base pairs spanned including gaps, whereas the length column for the total sequence row is the total number of base pairs in these scaffolds. The number column for placed pieces, for example, rocks, is the number of unitigs of that kind placed in the big scaffolds, whereas the length column gives the amount of sequence covered by that type but not by unitigs of the category above it, for example, 0.992 Mbp of the sequence for the joint data set was covered by rocks but not U-unitigs. Negative gaps are those where the assembler estimates that the two adjacent contigs should overlap but could not find one within the placement dictated by the bundles (*40*).

| | Joint | | WGS | | 6.5× WGS | |
|---|---|---|---|---|---|---|
| | Number | Length (Mbp) | Number | Length (Mbp) | Number | Length (Mbp) |
| *Scaffolds* | 25 | 116.176 | 26 | 116.306 | 43 | 114.348 |
| *Total gaps* | 1,434 | 2.030 | 1,887 | 2.322 | 7,111 | 5.790 |
| 100–150 kbp | 3 | 0.343 | 3 | 0.354 | 4 | 0.489 |
| 50–100 kbp | 0 | 0.000 | 1 | 0.097 | 6 | 0.430 |
| 10–50 kbp | 8 | 0.184 | 10 | 0.209 | 28 | 0.517 |
| 2–10 kbp | 237 | 1.283 | 245 | 1.371 | 649 | 2.636 |
| 0–2 kbp | 812 | 0.219 | 1,132 | 0.290 | 4,394 | 1.715 |
| Negative | 374 | – | 496 | – | 2,030 | – |
| *Total sequence* | 37,225 | 114.146 | 34,184 | 113.983 | 23,890 | 108.557 |
| U-unitigs | 7,446 | 110.581 | 7,164 | 110.604 | 8,007 | 103.933 |
| Rocks | 2,056 | 0.992 | 1,787 | 0.927 | 1,950 | 2.683 |
| Stones | 139 | 0.121 | 132 | 0.118 | 98 | 0.129 |
| Pebbles | 27,584 | 2.450 | 25,101 | 2.332 | 13,835 | 1.809 |

The level of assembly of the scaffolds larger than 100 kbp for the joint and WGS data sets, and a 6.5× WGS data set are compared in Table 2. The scaffolded regions for joint and WGS data sets were identical except that one 16.34-Mbp scaffold in the joint data set split into 10.45-Mbp and 5.64-Mbp scaffolds. There were 446 fewer gaps (23%) in the joint assembly, but these gaps constituted only 163 kbp (0.13%) of additional sequence, confirming that the additional coverage of the external data had a positive but small impact. Note carefully that in the joint assembly no advantage has been taken of the known relations between the shredded reads from a finished BAC and the relative proximity of draft reads from a given clone, thus it should not be surprising that the differences are small. We have not yet made design changes to the assembler to take advantage of this information. For example, of the 1434 gaps in the large scaffolds of the joint data set, 140 are spanned by finished BAC and P1 sequences that the assembler could have potentially joined.

The 6.5× WGS data set was produced by randomly sampling a 1-to-1 mix of 2-kbp to 10-kbp Celera reads totaling 6.5× coverage, in which 70% of reads were pairs and all of the BAC mates (13×) were included. For this data set, the assembler produced 43 scaffolds that are slightly contracted and fragmented versions of the 25 large scaffolds in the bigger data sets, containing more than 95% of their sequence. This confirms our earlier claims that one has a robust picture of a genome at 6.5× coverage with a whole-genome approach.

To evaluate the causes of the 1434 intra-scaffold gaps among the 25 large scaffolds of the joint data set, we examined the sequence adjacent to each gap to see if there were any reads in the data set overlapping into the gap and whether the end of the sequence was screened as being repetitive. A total of 927 of the gaps have no overlapping sequence at either end and are almost certainly sequencing gaps as confirmed by their generally small size. Another 244 have a matching screen item at both boundaries and are thus almost certainly unresolved repeats. Of the remainder, 164 appear to involve a sequencing gap and 99 appear to involve a repetitive element by virtue of having no overlap or a screen item at one end, respectively.

The assembly of a shotgun data set is not the last step in producing a genome; a finishing phase is necessary in which a certain level of gap closure by experimental

means is required. One laboratory working on a BAC-by-BAC project reported that for an average BAC size of 99 kbp sequenced at 8.57× coverage, there were an average of 3.8 gaps that required some directed sequencing implying an average contig size of 26 kbp (*27*). For all firm scaffolds of the joint assembly, the average contig size is 50.0 kbp, implying the equivalent finishing effort of 2.0 gaps per 99 kbp of BAC. However, although the distribution of the sizes of sequencing gaps is the same in the two scenarios, the WGA assembly has several hundred repeat-induced gaps that are generally of a larger size. Nonetheless, this comparison suggests that the total finishing effort for *Drosophila* may well end up being commensurate with a BAC-by-BAC approach.

## Validation of the *Drosophila* Assembly

*STS-level validation*. STS maps (*28*) for the chromosome arms were concatenated to give a whole-genome map that orders 2378 STSs, permitting comparison between this independent order and the WGS assembly. The STS sequences mapped a total of 114.8 Mbp of assembled sequence across 50 scaffolds to the *Drosophila* genome. There is excellent agreement between the STS order in the STS-content maps and the WGS assembly (Fig. 5). Twelve STSs were discarded from the study because they proved not to map to unique positions. Of the remaining 2366 sites, 2167 matched contigs in the assembly giving 2117 ordered pairs of STSs that could disagree between the two data sets (*29*). There were 17 ordering discrepancies, each of which was investigated. We have been able to localize nine of the discrepant STS sequences on the published clone-tiling path (CTP) (see below), the positions of which agree in each case to the Celera assembly position. Eight discrepancies are unresolved and remain under investigation.

*Clone-level validation and coverage*. The assembly of the WGS data set was compared to the finished and the 1.28× draft sequence available for the published CTP that covers most of the euchromatin of the genome (*30*). This allows us to identify the appropriate clone reagents for gap closure, and to verify the order and assembly of contigs in our scaffolds. As predicted from the results of the STS map comparison, the assembly is in excellent agreement with the published CTP (Fig. 6). There were only 11 discrepancies between the WGS assembly and the CTP. Each of these discrepancies was investigated and curated (*31*). One discrepancy is caused by a P1 clone on the tiling path that appears to be chimeric (*32*). The remaining 10 discrepancies were shown to be caused by placement errors in the CTP.

In an attempt to judge how much of the genome a pure whole-genome assembly captures, we compared the coverage of the 816 firm scaffolds of the WGS assembly to that of the CTP and associated sequence. This result is only indicative as it is difficult to precisely evaluate the intersection of our contigs with the light-shotgun data (*33*) because of repetitive sequence. The WGS assembly was estimated to miss approximately 2.99 Mbp of the sequence in clones of the CTP. Almost all of the missed sequence was present in reads not incorporated into firm scaffolds, and these absences were uniformly distributed across chromosomes, suggesting that this number estimates the amount of sequence in the larger gaps of the WGS assembly. In the converse direction, approximately 15 Mbp of WGS data could not be matched to CTP data. Not surprisingly, most of this involved contigs mapped to chromosome X and a region of 3L where the CTP is incomplete. From these numbers one might estimate that 105 Mbp of *Drosophila* are in the current physical map, and the WGS assembly has 3 Mbp of that in gaps, for a total of 97.1% coverage of the current physical map. One could then carry that number forward as an estimate of the percent of the euchromatin within the WGS assembly.

*Sequence-level validation*. A comparison of the complete published sequence of the 2.9-Mbp Adh region (*34*) against the 23 Celera contigs from the WGS assembly that cover it is shown in Fig. 7. We chose the Adh region because it was the longest contiguous stretch of finished sequence available. There are two levels of discrepancy—small point variations involving one, two, or three bases, and larger block variations involving from 33 bp to 9 kbp. All of the large variations are in our solutions to repeats, and we discuss them first.

There are 15 block-level differences between the two sequences, totaling 25 kbp. Three are Hobo-elements in our sequence that are strongly supported by the assembly and thus appear to be genuine repeat-level polymorphisms. Four are variations in the copy number of tandem duplications, where three are manually correctable overcompressions by one repeat unit in our sequence. The remaining eight block discrepancies are in the interiors of retrotransposons and appear to be due to incorrect pebble walks as described earlier. All involve on the order of 30- to 100-bp blocks with the exception of one substitution of 3500 bp and another retrotransposon that appears to be rearranged with respect to its long terminal repeats. There is thus room for improvement in the pebble repeat resolution phase, during which we did not adequately take advantage of interpebble mate pairs. All these discrepancies occur in regions covered solely by pebble-placed unitigs, and these constitute only 2.45% of the reconstruction. Altogether, we measured 9.5% of our *Drosophila* assembly as being

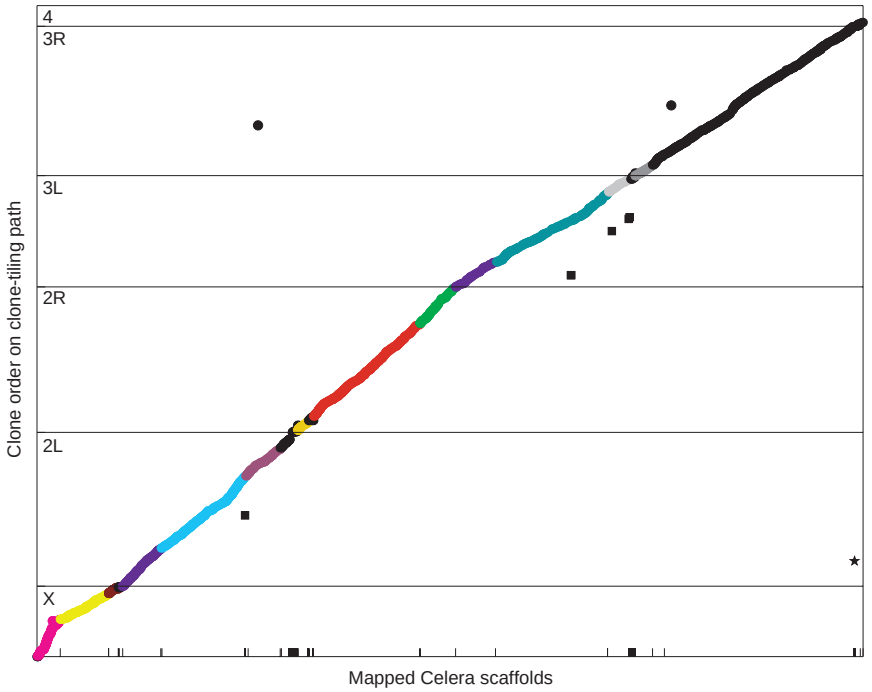


**Fig. 6.** Clone-tiling path (CTP) map. All mapped Celera scaffolds, oriented and ordered by both the STS-content map and the CTP sequence were plotted against all BAC/P1/cosmid clones ordered as they appear on the CTP. All "mutually unique regions" (*39*) between a clone and a contig are aggregated and displayed. The observed chimeric region (*34*) is marked by a star; evident misorderings in the tiling path are marked with square; repeat-induced "hits" are marked with a circle.

repetitive sequence, so the better part of most repeat constructions should prove correct with some variations in the interior of longer elements like those just described.

The concentration of individual base-pair discrepancies varies depending on whether the sequence is repetitive or not. The discrepancy in the repetitive regions is roughly 0.38%, whereas in the nonrepetitive sequence there are 140 differences for 0.0049% of the total. An examination of the differences in the nonrepetitive region indicates that 78 are in deep coverage regions of the assembly, where multiple alignments confirm our sequence. Therefore, one can bound our error rate in the nonrepetitive sequence as 62 in 2.82 Mbp or less than 0.0022%. The higher discrepancy rate in the repetitive region is explained by the use of unitig pebbles that are overcollapsed. Further details of the comparison are given in the legend. We thus project that we have a very high quality, ordered and mapped reconstruction of the nonrepetitive genome, with a draft-quality facsimile of the repetitive elements interspersed therein.

To get a broader picture of sequence quality, we scrutinized the results of BLAST searches of the WGS assembly against 104 high-quality, finished P1 or BAC clones, totaling 10.2 Mbp of sequence. After curating conflicts (*35*), we tabulated all discrepancies in high-scoring segment pairs (HSPs) longer than 10 kbp in both repetitive and unique regions, finding 63 inserts, 142 deletions, and 177 substitutions in 182.7 kbp of known repetitive sequence (0.021%), and 244 inserts, 182 deletions, and 231 substitutions in the remaining 7.085 Mbp of unique sequence (0.0092%). Of the sequence not in large HSPs, 77 kbp is simply clone sequence that is in gaps between contigs of the WGS assembly. There then remains 48 kbp of non-HSP sequence, 31 kbp of which is in known repeats and 17 kbp of which will likely be discovered to be either repeat polymorphisms or overcollapsed, unannotated tandem

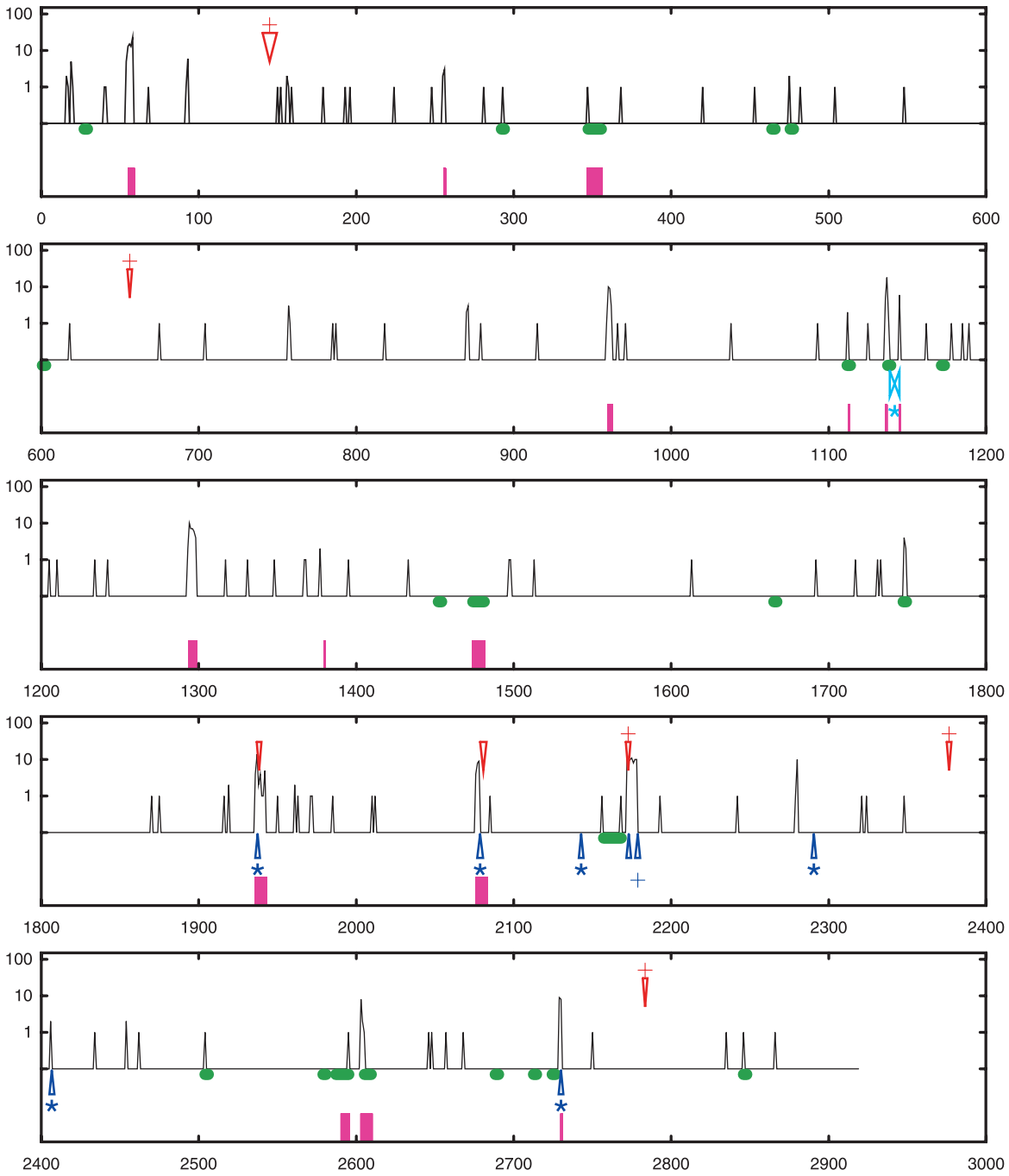


**Fig. 7.** Detailed comparison between the Adh region and WGS assembly. The *x* axis gives location (in thousands of base pairs) relative to the public Adh sequence. Peaks indicate the numbers of single-base mismatches between the two sequences in windows of length 1000 (log scale, or zero if there was perfect agreement). Purple boxes denote transposable elements in the public Adh sequence. Larger-scale discrepancies are as follows: green lozenges, gaps between Celera contigs; red inverted triangle, regions in the public Adh sequence that are absent from the Celera sequence ("deletions"); purple triangle, regions in the Celera sequence that are absent from the public sequence ("insertions"); cyan X, inversion of a region of one sequence relative to the other. A star associated with an insertion or inversion indicates the presence of transposable elements. A plus sign indicates that the associated insertion or deletion involves tandem duplications.

repeats. This amount of this unaccounted for sequence is proportionally less than that of the block-level differences in the more detailed comparison against the Adh region. The methodology here is limited, in that poor repeat constructions will not occur in long HSPs and thus cause an undercount of individual base differences in repeat reconstructions, and the sequence is elsewhere being underannotated as repeat, necessarily overcounting individual base differences in nonrepetitive sequence. However, it does support an extrapolation of the precise results given for the 2.5% of the genome in the Adh region. An initial comparison between our results and the 30 Mbp of finished *Drosophila* sequence has also been performed (*36*).

*Finishing report*. Gap closure has been a collaboration between Celera and the BDGP sequencing groups at Lawrence Berkeley National Laboratory and Baylor College of Medicine (*11*). Of the total of 1630 gaps in the joint assembly data set, 12 gaps have been closed by finished BAC sequence, 17 negative gaps have been closed by computation, and another 302 gaps have been closed by the BDGP via directed gap-filling. The average size of the successfully closed gaps is 771 bp; the estimated size of the remaining gaps is 2120 bp. There are 1299 gaps currently remaining in the assembly, and at the current rate of closure, we could reach fewer than 100 gaps remaining in the euchromatic portion of the genome by the end of July 2000.

*What remains*. Analysis of the results of our assemblies is ongoing. In particular, (i) we continue to work on mapping assembled contigs near the centromeres; (ii) we continue to monitor for the possible presence of large, duplicated regions; (iii) a detailed comparison, as for the Adh region, between our results and the 30 Mbp of finished *Drosophila* sequence is under way; and (iv) further internal consistency checks on assemblies are contemplated. Analysis updates will occur periodically and will be made available on the World Wide Web (*36*). As of this writing, we have an assembly of *Drosophila* suitable for a wide range of biological studies. We continue to work on improvements for repeat resolution in order to consistently achieve a quality of sequence in these regions that is closer to community standards for finished sequence. We believe that with sufficient time, the several algorithmic avenues we are exploring will yield such improvements. In the interim, we felt compelled to release the assembly at its current standard because of its value to the scientific community.

**References and Notes**


1. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 12 (1977).
2. F. R. Blattner *et al.*, *Science* **277**, 1453 (1997).
3. H. W. Mewes *et al.*, *Nature* **387** (6632 suppl.), 7 (1997).
4. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
5. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
6. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, *J. Mol. Biol.* **162**, 4 (1982).
7. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
8. J. Weber and H. Myers, *Genome Res.* **7**, 401 (1997).
9. P. Green, *Genome Res.* **7**, 410 (1997).
10. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
11. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
12. G. Rubin *et al.*, *Science* **287**, 2204 (2000).
13. In addition to judicious algorithmic application of mate pairs, we are fortunate because with the introduction of capillary gel sequencers, the primary source of false-pairing information for end reads disappears, as a sample is now forced to flow down a tube as opposed to meandering over a slab gel. With careful robotics and library construction the false-pairing rate on mate pairs can be kept to less than 1%.
14. The ES40 utilizes a 667-MHz Alpha 21264a processor running Tru64 UNIX. Each CPU receives a score of 413 and 500, respectively, for the integer and floating point SPEC 2000 benchmarks (see www.spec.org).
15. During the first 9 months of development, when no significant amount of real data was available, we used a simulator called celsim [E. W. Myers, in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, Heidelberg, Germany, August 1999; T. Lengauer *et al.*, Eds. (American Association for Artificial Intelligence, Menlo Park, CA, 1999), pp. 202–210] that could either take a mosaic of known sequence, for example, *Caenorhabditis elegans* or *Saccharomyces cerevisiae*, and simulate the shotgun process on it or generate synthetic DNA with controllable repeat characteristics and simulate a shotgun of it.
16. Starting with the raw sequencing data, one generally trims off a prefix and suffix of a read that is too inaccurate. In the early days, when assembly software was not very robust, one trimmed aggressively to be certain of a having a highly accurate remainder. But longer reads imply less oversampling and hence greater efficiency, so as software improved, the cutoff was relaxed to the point where today many use and report read lengths for trimming at the 90 to 95% accuracy level. One must return to a tight 98% stringency for whole-genome shotgun sequencing in order to avoid false overlaps.
17. We used a software package developed by Paracel, Inc., building on the ideas originally published by B. Ewing, L. Hillier, M. C. Wendi, and P. Green [*Genome Res.* **8**, 175 (1998)].
18. The accuracy of reads was evaluated by finding all reads that were sampled from the 29 Mbp of finished *Drosophila* sequence produced by the BDGP and EDGP. Comparison of the trimmed portions of such reads against the finished sequence was used to determine the accuracy of the read.
19. The largest source of error in pairing, lane tracking error, disappears with capillary gel sequencing. The remaining sources of error are chimerism in the insert library and sample tracking in the lab. One of the significant advantages of a whole genome approach is that in concept only three libraries are needed, so one can very carefully craft and assure the quality of these libraries. We estimated the chimerism rate of the library at 0.01% and required the laboratory protocols to be of sufficient quality that a plate would correctly track through the entire sequencing pipeline at least 99.5% of the time. The actual false-pairing rate was measured by examining all pairs wherein one read could uniquely be localized to the interior portion of a finished BAC sequence of *Drosophila*.
20. R. A. Hoskins *et al.*, *Science* **287**, 2271 (2000).
21. A perfect shredding of a sequence is a set of reads covering the sequence that (i) all have the same length; and (ii) overlap the read that immediately follows them by exactly the same amount.
22. S. Altschul, W. Gish, W. Miller, E. W. Myers, D. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
23. E. W. Myers, *J. Comp. Biol.* **2**, 2 (1995).
24. Suppose there are $F$ fragments in a database and the genome size is estimated as $G$. For a unitig with $k$ fragments and distance $\rho$ between the start of its first fragment and the start of its last fragment, the probability of seeing the $k - 1$ start points in the interval of length $\rho$, given the unitig is not oversampled, is $[(\rho F/G)^k/k!]\exp(-\rho F/G)$. If the unitig was the result of collapsing two repeats, then the probability is $[(2\rho F/G)^k/k!]\exp(-2\rho F/G)$. The log of the ratio of these two probabilities, $(\log e)\rho F/G - (\log 2)k$, is our A statistic.
25. G. Churchill and M. Waterman, *Genomics* **14**, 89 (1982).
26. The computer program continues to be refined. The computer algorithms used and the correct version described here are the subject matter of a pending patent application. We are open to collaborations involving this software under terms beneficial to all parties.
27. S. Salzburg, The Institute for Genome Research, personal communication. The data were taken from 10 randomly sampled BACs from *Arabidopsis thaliana*.
28. Because of the variable status of the STS map across chromosomes, several protocols were required to find these sequences and to compare them against the WGS assembly. For the X chromosome, we used public sequences from cosmid fragments wherever there was complete sequence available. Sequence tags were identified by searching for markers within the cytogenetic regions included on the X chromosome in Flybase, and BAC end sequences mapped to the X and STS-content map contigs from Berkeley, ordered only by their reported cytological range (see www.fruitfly.org and http://flybase.bio.indiana.edu). For chromosomes 2 and 3, we used the 40-bp overgo sequences from the Berkeley STS-content map, which were ordered by BDGP against panels of BACs (*20*). For chromosomes 4 and Y, we used publicly available sequences ordered based on their cytology, including genes and one set of cosmid ends (see www.fruitfly.org and flybase.bio.indiana.edu). BLAST (*38*) was used to locate the sequence tags in the assembly, using 95% identity and length of 99 bp as our cutoff for tags on chromosomes X, 4, and Y; the length cutoff was reduced to 38 bp for the 40-bp overgo sequences on chromosomes 2 and 3. We regarded STSs hitting the assembly many times as unreliable for sequence localization for the purposes of this study, and such sequences were eliminated from consideration. When a scaffold showed an inconsistent map association, the location of the STS sequence was then checked against the sequence in the clone-tiling path map.
29. A total of 199 STSs was not located on the assembly, 57% of which are on the X chromosome.
30. There are a total of 1135 clones in the path: 380 were finished and 755 had been light-shotgunned at 1 to 1.5× sequence coverage. For the draft sequence, we used the assembled sequences submitted to GenBank (www.ncbi.nlm.nih.gov/Entrez/batch.html). Finished and light-shotgun clones were treated separately; comparisons of assembled contigs to finished clones were made directly, whereas those to light-shotgun clones were made by comparison with a conglomerate of the unassembled shotgun fragments. BLAST (*38*) was used to search all of the WGS placed contigs against all light-shotgun and finished clones in the CTP. An E-value cutoff of $10^{-30}$ and identity cutoff values of 99% for finished data and 95% for light-shotgun data were used in the BLAST comparisons.
31. To resolve discrepancies, the validity of both orders was examined by sequence comparison between the discrepant clone and assembly region and all other clone sequences in the tiling path. Further, the clone and assembly regions in question were compared with the STS-content map. If the weight of evidence data supported the order of either the assembly or the tiling path, we concluded that the supported order was correct.
32. The first 30 kbp of tiling clone DS08493 does not match the Celera contig covering the entire region of the tiling path. The adjacent tiling clone BACR11E09 does not match DS08493 across the chimeric junction; instead it agrees with the Celera contig. The GenBank accession number of the clone in question is AC004422.
33. We identified 234 clones that are entirely covered by WGS assembled contigs; no clone is completely missed by Celera contigs. For clones that were partially covered, clone regions of at least 50 bp (excluding place-

holder Ns) that have no BLAST hit at 99% identity for finished data and 95% identity for light-shotgun data were considered uncovered. The percentage of each clone not hit by Celera sequence was calculated by dividing the total length of the uncovered sequence by the sequence length of the clone. The total number of nucleotides that have no coverage in the Celera assembled contigs was calculated by summing the regions of no hits for all the clones that covered Celera contigs by less than 90% (95% for finished clones). This cutoff value was chosen to eliminate the occasionally low quality of sequences in the clone sequence data. The cutoff value of 90% was determined by the amount of no-hit sequences in 16 light-shotgun clones that are fully contained within three Celera contigs. A higher cutoff value (95%) was used for the finished data than for the light-shotgun data, because finished clones have better sequence quality. The total amount of uncovered sequence for each light-shotgun clones was calculated by multiplying the no-hit percentage of the clone by the clone length as determined by sizing on agarose gels (*36*). For those light-shotgun clones with unreported insert sizes, the sequence length, excluding Ns, was used instead. For finished clones, the amount of uncovered sequence was calculated by multiplying the no-hit percent of the clone by the clone's length. We created 7-kbp subcontig blocks and considered each block to be fully present in the draft sequence if it was hit by at least 500 bp of external sequences. We chose these parameters conservatively, based on the fact that at 1× sequence coverage, the chance of failing to sample a 7-kbp region covered by a light-shotgun clone is 1 in $10^6$. For the WGS assembly, we identified 1380 blocks that were hit by less than 500 bp of clone sequence and 794 blocks that were completely missed by the clone sequence. The total number of missed blocks is 2174, which represents a total 15.2 Mbp.

34. M. Ashburner *et al.*, *Genetics* **153**, 179 (1999).
35. Seven conflicts were identified in this study, six of which appear to be owing to transposable elements. The remaining represents a 30-kbp insert within a Celera contig that does not match the corresponding clone. This discrepancy is still under investigation.
36. www.sciencemag.org/feature/data/1049666.shl
37. S. Altschul *et al.*, *Nucleic Acids. Res.* **25**, 3389 (1997).
38. R. A. Hoskins, personal communication.
39. In order to align the Celera sequences unambiguously to the external data, all significant HSPs at the parameters given in (*27*) were screened to identify "mutually unique regions" where the clone and contig sequences have a unique, reciprocal match relation.
40. Most negative gaps arise because of inaccuracies in the distances implied by bundles—the bundle implies a small amount of overlap between two contigs because it is actually short, whereas the reality is that there is a small gap at that location. In a very small number of cases, there is an overlap, but it is because the distance estimate is too long by 3 standard deviations, or because there is a small bit of foreign DNA at the tip of a contig because of untrimmed vector or a chimeric read. None of these negative gaps has yet been found to imply incorrect assembly.
41. We wish to thank H. Smith and S. Salzburg for the many collegial exchanges, M. Peterson and his team for keeping the machines humming, R. Thompson and his staff for providing us with an environment conducive to such an intense effort, and A. Glodek, C. Kraft, and A. Deslattes Mays, and their staff for getting the data to us.

REVIEW

# Comparative Genomics of the Eukaryotes

**Gerald M. Rubin,[1] Mark D. Yandell,[3] Jennifer R. Wortman,[3] George L. Gabor Miklos,[4] Catherine R. Nelson,[2] Iswar K. Hariharan,[5] Mark E. Fortini,[6] Peter W. Li,[3] Rolf Apweiler,[7] Wolfgang Fleischmann,[7] J. Michael Cherry,[8] Steven Henikoff,[9] Marian P. Skupski,[3] Sima Misra,[2] Michael Ashburner,[7] Ewan Birney,[7] Mark S. Boguski,[10] Thomas Brody,[11] Peter Brokstein,[2] Susan E. Celniker,[12] Stephen A. Chervitz,[13] David Coates,[14] Anibal Cravchik,[3] Andrei Gabrielian,[3] Richard F. Galle,[12] William M. Gelbart,[15] Reed A. George,[12] Lawrence S. B. Goldstein,[16] Fangcheng Gong,[3] Ping Guan,[3] Nomi L. Harris,[12] Bruce A. Hay,[17] Roger A. Hoskins,[12] Jiayin Li,[3] Zhenya Li,[3] Richard O. Hynes,[18] S. J. M. Jones,[19] Peter M. Kuehl,[20] Bruno Lemaitre,[21] J. Troy Littleton,[22] Deborah K. Morrison,[23] Chris Mungall,[12] Patrick H. O'Farrell,[24] Oxana K. Pickeral,[10] Chris Shue,[3] Leslie B. Vosshall,[25] Jiong Zhang,[10] Qi Zhao,[3] Xiangqun H. Zheng,[3] Fei Zhong,[3] Wenyan Zhong,[3] Richard Gibbs,[26] J. Craig Venter,[3] Mark D. Adams,[3] Suzanna Lewis[2]**

A comparative analysis of the genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*—and the proteins they are predicted to encode—was undertaken in the context of cellular, developmental, and evolutionary processes. The nonredundant protein sets of flies and worms are similar in size and are only twice that of yeast, but different gene families are expanded in each genome, and the multidomain proteins and signaling pathways of the fly and worm are far more complex than those of yeast. The fly has orthologs to 177 of the 289 human disease genes examined and provides the foundation for rapid analysis of some of the basic processes involved in human disease.

With the full genomic sequence of three major model organisms now available, much of our knowledge about the evolutionary basis of cellular and developmental processes will derive from comparisons between protein domains, intracellular networks, and cell-cell interactions in different phyla. In this paper, we begin a comparison of *D. melanogaster, C. elegans*, and *S. cerevisiae*. We first ask how many distinct protein families each genome encodes, how the genes encoding these protein families are distributed in each genome, and how many genes are shared among flies, worms, yeast, and mammals. Next we describe the composition and organization of protein domains within the proteomes of fly, worm, and yeast and examine the representation in each genome of a subset of genes that have been directly implicated as causative agents of human disease. Then we compare some fundamental cellular and developmental processes: the cell cycle, cell structure, cell adhesion, cell signaling, apoptosis, neuronal signaling, and the immune system. In each case, we present a summary of what we have learned from the sequence of the fly genome and how the components that carry out these processes differ in other organisms. We end by presenting some observations on what we have learned, the obvious questions that remain, and how knowledge of the sequence of the *Drosophila* genome will help us approach new areas of inquiry.

## The "Core Proteome"

How many distinct protein families are encoded in the genomes of *D. melanogaster*, *C. elegans*, and *S. cerevisiae* (*1*), and how do these genomes compare with that of a simple prokaryote, *Haemophilus influenzae*? We carried out an "all-against-all" comparison of protein sequences encoded by each genome using algorithms that aim to differentiate paralogs—highly similar proteins that occur in the same genome—from proteins that are uniquely represented (Table 1). Counting each set of paralogs as a unit reveals the "core proteome": the number of distinct protein families in each organism. This operational definition does not include posttranslationally modifed forms of a protein or isoforms arising from alternate splicing.

In *Haemophilus*, there are 1709 protein coding sequences, 1247 of which have no sequence relatives within *Haemophilus* (*2*). There are 178 families that have two or more paralogs, yielding a core proteome of 1425. In yeast, there are 6241 predicted proteins and a core proteome of 4383 proteins. The fly and worm have 13,601 and 18,424 (*3*) predicted protein-coding genes, and their core proteomes consist of 8065 and 9453 proteins, respectively. It is remarkable that *Drosophila*, a complex metazoan, has a core proteome only twice the size of that of yeast. Furthermore, despite the large differences between fly and worm in terms of development and morphology, they use a core proteome of similar size.