BAC library construction and BAC end sequencing of five *Drosophila* species: the comparative map with the *D. melanogaster* genome

Katsuhiko Murakami^{1,2}, Atsushi Toyoda¹, Masahira Hattori¹, Yoko Kuroki¹,
Asao Fujiyama¹, Toshio Kojima¹, Muneo Matsuda³, Yoshiyuki Sakaki¹
and Masa-Toshi Yamamoto^{4*}

¹RIKEN Genomic Sciences Center, 1-7-22, Suehiro-cho, Tsurumi, Yokohama, Kanagawa
230-0045, Japan

²Current address: Integrated Database Group, Japan Biological Information Research
Center (JBIRC), Japan Biological Informatics Consortium,
Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan

³Department of Biology, School of Medicine, Kyorin University, 6-20-2 Shinkawa,
Mitaka, Tokyo 181-8611, Japan

⁴*Drosophila* Genetic Resource Center, Kyoto Institute of Technology, Saga
Ippongi-cho, Ukyo-ku, Kyoto 616-8354, Japan

(Received 10 July 2007, accepted 11 March 2008)

We constructed and characterized arrayed bacterial artificial chromosome (BAC) libraries of five *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. auraria*, and *D. ananassae*), which are genetically well characterized in the studies of meiosis, evolution, population genetics, and developmental biology. The BAC libraries comprise 8,000 to 12,500 clones for each species, estimated to cover the most of the genomes. We sequenced both ends of most of these BAC clones with a success rate of 91%. Of these, 53,701 clones consisting of non-repetitive BAC end sequences (BESs) were mapped with reference of the public *D. melanogaster* genome sequences. The BES mapping estimated that the BAC libraries of *D. auraria* and *D. ananassae* covered 47% and 57% of the *D. melanogaster* genome, respectively, and those of *D. melanogaster*, *D. sechellia*, and *D. simulans* covered 94–97%. The low coverage by BESs of *D. auraria* and *D. ananassae* may be due to the high sequence divergence with *D. melanogaster*. From the comparative BES mapping, 111 possible breakpoints of chromosomal rearrangements were identified in these four species. The breakpoints of the major chromosome rearrangement between *D. simulans* and *D. melanogaster* on the third chromosome were determined within 20 kb in 84E and 30 kb in 93E/F. Corresponding breakpoints were also identified in *D. sechellia*. The BAC clones described here will be an important addition to the *Drosophila* genomic resources.

Key words: BAC end sequence (BES), BAC library, chromosomal rearrangement, *Drosophila*, genome evolution

INTRODUCTION

Since *D. melanogaster* genome sequence was published in 2000 (Adams et al., 2000), *D. pseudoobscura* genome sequence was released in 2003, and additional ten *Drosophila* species genome sequences were reported recently (Clark et al., 2007; Gilbert, 2007). The genomes of the *melanogaster* subgroup, *D. erecta*, *D. sechellia*, *D. simulans* and *D. yakuba*, were included. Sequencing and

comparative analysis among the genomes of evolutionarily related species are important approach which will give us great insights into the functional aspects of genomes, for example, genome annotation and identification of genome features, as discussed in several reports (Bergman et al., 2002; Richards et al., 2005; Zdobnov et al., 2002).

BAC libraries are the important and basic resource for any kind of genomic analyses concerning the whole genome, such as full sequencing of BAC inserts, construction of BAC and BAC-based maps (Ren et al., 2003), com-

Edited by Etsuko Matsuura

* Corresponding author. E-mail: yamamoto@kit.jp

parative analysis of gene structures and synteny (Larkin et al., 2003; Leeb et al., 2006), and powerful transformation technologies (Bischof et al., 2007; Oberstein et al., 2005; Venken et al., 2006) using DNA fragments larger than 100 kb to specific sites in the *Drosophila* genome.

We thus constructed BAC libraries and analysed the BAC end sequences (BESs) of five species, which were *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. auraria*, and *D. ananassae*. *D. melanogaster* was selected as a control in order to measure the various qualities of the BAC libraries and the construction processes. The other four species are extensively utilized and recognized important organisms for research on speciation, hybrid incompatibility (Sawamura et al., 1993, 2004), food preference (Higa and Fuyama, 1993), mating behaviour (Tomaru et al., 1995), and male recombination (Tobari, 1993). *D. auraria* was specially selected for a species in the *montium* subgroup, which are interesting model organism for the research on cold resistance and diapause in adaptation biology (Kimura, 1988). The five species are abbreviated by DME1 for *D. melanogaster*, DSM1 for *D. simulans*, DSE1 for *D. sechellia*, DAU1 for *D. auraria*, and DAN1 for *D. ananassae*. All five species belong to the *melanogaster* species group. As these species diverged at different times as shown in Fig. 1 (Tamura et al., 2004), the BAC and the BES-based comparison between these species would be expected to provide sufficient information and abundant molecular resources for various research with the references of the genome sequence information recently reported (Clark et al., 2007; Gilbert, 2007).

After sequencing BESs, we assessed the sequence quality in order to perform accurate BES mapping. The

high quality portion of the sequences were then compared with the *D. melanogaster* genome sequence (Berkeley Drosophila Genome Project (BDGP) release 3: <http://www.fruitfly.org/>). These mapping processes allowed us to identify which BAC clone includes the homologous genes or regions corresponding to the region of interest in *D. melanogaster*. The high quality BESs data have been deposited to DDBJ under accession nos. AG909574 through AG981700. All information concerning the BAC libraries and BESs are also available via Drosophila Genetic Resource Center, Kyoto Institute of Technology (<http://www.dgrc.kit.ac.jp/en/>), and searchable at the web site (<http://fly.gsc.riken.jp/>). The results for genome-wide comparative BES mapping can be viewed through the web site of RIKEN with data embedding into UCSC genome browser (Kent et al., 2002).

MATERIALS AND METHODS

Drosophila strains *Drosophila* strains from which genomic DNA was extracted were Hikone-R of *D. melanogaster*, Rakujuen RM13 of *D. simulans*, ss01 of *D. sechellia*, A662 of *D. auraria*, and AABBg1 of *D. ananassae*. All strains are available from DGRC (Kyoto Inst. Technology), Kyorin University or Ehime University through the web sites (<http://kyotofly.kit.jp/stocks/>).

Construction of BAC libraries BAC libraries were constructed according to the procedures previously described (Fujiyama et al., 2002). In brief, frozen fly bodies were homogenized in 0.2 M sorbitol, 20 mM Tris-HCl (pH 7.5), and 10 mM EDTA, then centrifuged at 5,000 rpm for 10 min after removal of large debris. Sediments were embedded in 1% agarose gel, treated with *Sac* I, and subjected to pulsed-field gel electrophoresis. The DNA fragments ranging from 80 to 200 kb were isolated and ligated with pKS145 vector. Transformation was carried out electronically using *E. coli* DH10B as a host strain. Ampicillin-resistant transformants were collected and stored in 384-format plates.

Sequencing of BAC ends BAC clones from each library were rearranged into 96 well plates and then cultured overnight in 1.5 ml of 2x LB medium containing 50 µg/ml ampicillin at 37°C. The BAC DNA was isolated by using a Montage BAC 96 Miniprep kit (Millipore) and dissolved in 35 µl of 10 mM Tris-HCl buffer (pH 8.0). Cycle sequencing reaction was performed in a 15 µl reaction volume containing 10 µl BAC DNA solution, 1.5 µl Big Dye terminator Ready Reaction mix (Applied Biosystems) and 4.8 pmole of sequence primer. Custom-designed sequence primers at the forward and reverse sides were TTCCCAGTCACGACGTTGTA and TCAC-TATAGGGCGGCCGCTA, respectively. PCR conditions were as follows: one cycle of 5 min at 95°C, followed by 75

Fig. 1. Phylogenetic relationships of the *melanogaster* Species. The divergence times are based on the work in Tamura et al. (2004). The species for which we constructed BAC libraries are written in bold.

cycles of 30 sec at 95°C, 10 sec at 50°C and 4 min at 60°C. The reaction mixture was purified by isopropanol precipitation followed by 70% ethanol wash and loaded on ABI 3700 automated capillary DNA sequencers (Applied Biosystems).

Sequence Processing Base-calling and vector identification was performed by the Phred program (Ewing and Green, 1998; Ewing et al., 1998), which assigns a quality value (QV) to each base. After the Phred scoring step, we further eliminated vector sequences containing *E. coli*-derived sequences from the sequence as follows. We queried each BES on both the *E. coli* and *D. melanogaster* genomes by BLASTN (Altschul, 1997). If a BES hit with *E. coli*, and the hit meets the criteria (length ≥ 120 bp and percent identity $\geq 90\%$), then the hit was regarded as a sequence derived from *E. coli*. We confirmed that this criterion does not eliminate the *D. melanogaster* sequence.

Filtering Bacteria Sequences We identified additional bacterial sequences, such as a commensal bacterium, *Wolbachia*, by searching with the BESs on the non-redundant nucleotide database, called 'nt' provided by NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). When significant hits (condition "length ≥ 50 bp and identity $\geq 80\%$ ") were observed against bacterial sequences, we regarded them as bacterial contamination and eliminated them from the BES data set.

Multi-block setting to determine high-quality regions In determination of the region of BESs with high enough score to be analyzed, we have developed a new approach by which unnecessary removal of useful sequence information is reduced as much as possible. Determination of a high QV region in a BES is often carried out by extracting a continuous region of QV ≥ 20 and length ≥ 100 bp. It is generally observed that QVs are low at the first 1–50 bp region and the end of the BES, therefore a QV-position curve usually shows a single broad peak, which shows a high QV region to be selected. However for some BESs, the QV-position curve drops under 20 (QV) around 200 bp position, which is possibly caused by a dye blob as pointed out previously (Zhao et al., 2001). The low quality regions typically are several bases in length. In the case that there are two or more high quality peaks in a BES, the single-peak assumption allows us to obtain only a single block of high QV region either before or after the short low QV regions, resulting in the elimination of other high QV regions from the BES. To solve the problem, and significantly extend the query sequence without lowering the mapping quality, we employed all long high QV regions as a query, even if the sequence contains intervening low QV regions, as long as they were short. To utilize this idea, we first identified blocks with high QV (QV ≥ 20 runs continuously at least

100 bp), and then we combine the blocks containing intervening low QV regions. We call the procedure 'multi-block setting'. This is useful especially for comparative BES mapping to identify homologous regions in a genome of related species. Because even homologous or syntenic DNA sequences of phylogenetically distant species sometimes differ significantly, (i.e. nucleotide identity $< 75\%$), it is important to use a longer sequence as the query for the accurate identification of homologous regions. The low frequency of low quality or incorrect bases in a query DNA sequence may be a minor problem when the query is long enough. For instance, in the case that a BES contains a 10 bp region with the QV = 10, the error is estimated to be only 1 base. As the number of sequence errors in the low quality regions combined by the method should be very small, the sequence quality hardly gets worse by the multi-block setting. The multi-block settings extended the BES by 46%–66% longer than before, recovering those useful regions. The average extended regions per BES are ranged from 178 bp to 218 bp, while low QV bases incidentally included in the query were only 1.9%–2.8% of the resultant BES on average.

Self Comparison For self-comparison, BLASTN was performed with the parameters '-U T -F "m D" -e 1e-4'. The "U" option uses lower-case filtering of FASTA sequence. The "W" option specifies the word size in initial seeds. The "F" option uses filtering by the *Dust* program. The "e" option specifies expectation value, or the expected number of sequences to be reported. We set a criterion to identify the region (assumed to be novel repetitive sequence) in the BES which has multiple hits in the self comparison as follows. When only one region of BES matched with only another BES, it might be the same location of the genome was sequenced as the two BESs. To be more confident with the interpretation that the matches are novel repeat sequences, we required two BES matches (i.e. three similar sequences) to regard them as repeat sequences. We describe the conditions as below more rigorously. When a region (from x_1 to x_2 in base position, $x_1 < x_2$) in the query sequence hits with the region in the subject (another BES) sequence (from y_1 to y_2) with plus direction, then the hit is represented as $h(\vec{s}_1, \vec{s}_2)$, where, $\vec{s}_1 = (x_1, y_1)$, $\vec{s}_2 = (x_2, y_2)$. Here $\vec{s}_1$ is a vector to represent the first base of the hit, and $\vec{s}_2$ is a vector to represent the last base of the hit. Let us suppose there is another hit $h(\vec{t}_1, \vec{t}_2)$ for the same query sequence but with another BES with the same (plus) direction, where $\vec{t}_1 = (x'_1, y'_1)$, $\vec{t}_2 = (x'_2, y'_2)$, and $x'_1 < x'_2$, $y'_1 < y'_2$. The *region distance* is defined as $d = \sqrt{|x_1 - x'_1|^2 + |x_2 - x'_2|^2}$, which indicates the degree of difference between the positions of the edges of the two hits along the query. Let $x_{1,\min} = \min(x_1, x'_1)$, $x_{1,\max} = \max(x_1, x'_1)$, $x_{2,\min} = \min(x_2, x'_2)$, $x_{2,\max} = \max(x_2, x'_2)$, $U = \{x_{1,\min}, x_{1,\min} + 1, \dots, x_{2,\max}\}$, $V = \{x_{1,\max},$

$x_{1,\max}+1, \dots, x_{2,\min}\}$. Then, U is the union of the two hits in the query, and the V is the overlapped region between the two hits in the query. Then the *overlap rate* is defined as $r = |V|/|U|$. A set of selected hits were called *multi-hits* if all the hits have e-values ≤ 0.01 , and *region distance* d (described above) ≤ 20 bp and *overlap rate* r (described above) ≥ 0.5 . The multi-hits portions were masked from BESs data set before comparative BES mapping.

Comparative BES mapping BESs were mapped onto the *D. melanogaster* genome sequence (Release 3) taken from the UCSC WWW site (<http://genome.ucsc.edu/>, version of January 2003), which was provided by the Berkeley Drosophila Genome Project (BDGP). BLASTN was performed with the parameter set “-r 1 -q -1 -G 1 -E 2 -U T -F “m” -e 1e-2” (called ‘arranged set’), as well as the default parameter set. The parameters indicate match score, mismatch penalty, gap initial penalty, gap extension penalty, set lower cases non-seeds, filtration, and maximum e-value to report. The parameter was based on the suitable set for genome comparison where target identity of a match is around 75%, while the default parameters assume that the target identity of a match is about 90% (Korf et al., 2003). If the BLAST hits were split, a chaining rule was applied to select and combine consistent split hits for a query BES, as PIPMaker does (Schwartz et al., 2000). The BLAST hits with heterochromatic sequences were not considered.

Consensus construction of ‘self-match’ Matched regions were extracted and clustered by the *blastclust* program in BLAST suite with parameters “-p F -L 0.9 -S 99 -b F” (Korf et al., 2003). This clusters sequences that are expected to be nearly identical (99 percent identity, 90 percent coverage). For the top 30 clusters for each species, we aligned the sequences for each cluster by CLUSTALW version 1.83 (Thompson et al., 1994) and constructed a consensus by the *prophesy* program in EMBOSS package (Olson, 2002) for each cluster. Then BLASTN with default parameters was performed against the *nt* database (non-redundant nucleotide database in NCBI, <ftp://ftp.ncbi.nlm.nih.gov/blast/db/>).

Breakpoint prediction When a pair of BESs of a BAC clone were mapped to the genome with significant hits, the BESs or BAC clone was categorized as either *valid* or *invalid*, according to the following definition. If the two BES hits on the *D. melanogaster* genome are inward indirection and their distance L_{in} on the *D. melanogaster* genome is reasonable ($L_{in} \leq 300$ kbp), the BESs were called *valid*. In the ‘unmapped’ BESs, if each BES was mapped on the wrong direction or on different chromosomes or $L_{in} > 300$ kbp, the BESs were called *invalid*. The *invalid* BESs imply that the BESs are correlated with any chromosomal rearrangements, such as long insertion and dele-

tion, while *valid* BESs are not. The *valid* BESs for which the insert sequences overlap each other can be assembled to be a contig. The *invalid* BESs were clustered when their locations are close and the directions are the same. The difficulty arises when there are contradictions among *valid* and *invalid* BESs due to any artefacts, or shortness of BESs. There would be artefacts if there are duplications and subsequent mutations in any of the species, which make it difficult to identify real syntenic regions. To reduce artefacts, we selected the regions with breakpoint candidates around which we have more than one supports of both *valid* and *invalid* BESs.

RESULTS AND DISCUSSION

Sequence quality: Q20 length and success rates

We constructed BAC libraries composed of between 8,000 and 12,500 clones for the five species. The clone numbers of BAC libraries and BESs are shown in Table 1 (Initial total BAC libraries and Initial total BESs). Base-calling was performed by the Phred program (Ewing and Green, 1998; Ewing et al., 1998), which assigns a quality value (QV) to each base. When a base has a Phred QV ≥ 20 , the error rate is $\leq 1\%$. Such a base is called a “Q20 base.” To assess the quality of a sequence, the “Q20 length” was defined (Zhao et al., 2000, 2001). The Q20 length is the number of Q20 bases in each sequence. Note that it is not a real length but the count of high quality bases, though it has the term “length.” Figure 2 shows the cumulative Q20 length distribution. The Q20 length for other BAC end sequencing projects is around 400–500 bp (Larkin et al., 2003; Zhao et al., 2000, 2001).

Fig. 2. Q20 Length Distribution of Each Species. The X-axis represents the Q20 length of each sequence. The Y-axis represents the accumulated fractions of BAC end sequences of the species. The Q20 length of about half of all the BESs is about 600 bp. For the species DAN1, about the half of the sequences are relatively low quality (a little upper than the others). Only 10% of the sequences were low quality, when we set the threshold of Q20 length ≤ 100 bp ($X = 100$).

Table 1. Statistics of the BAC end sequences

Species	DME1	DSM1	DSE1	DAU1	DAN1	Total (Average)
Initial total BAC libraries	8,064	10,368	10,368	8,448	11,520	48,768
Initial total BESs	16,128	20,736	20,736	16,896	23,040	97,536
Success rate (%BES, success/total)	88.9	91.3	92.7	92.2	88.3	(91)
Average length (bp)	597	579	576	598	565	(583)
Variance (bp)	124	127	131	141	156	(136)
Total (bp)	8,561,716	10,981,493	11,085,630	9,333,109	11,513,759	51,475,707
BESs one-end	993	1,078	866	977	1,693	5,607
BESs mate-pairs	6,674	8,930	9,180	7,304	9,331	41,419
Q20-HQ BESs ^a	14,341	18,938	19,226	15,585	20,355	88,445
Q20-passed BESs ^b	13,042	18,024	17,944	13,780	14,375	77,165
Masked rate (%bp) ^b	16.0	11.4	15.8	12.0	10.8	(13.2)
Passed BESs by RM ^c	11,157	16,063	15,195	12,434	12,688	67,537
Masked (%bp) ^c	2.9	2.6	2.9	5.1	4.3	(3.6)
Passed BESs by MM ^d	10,682	15,354	14,373	10,766	10,092	61,267
Masked by MM (%bp) ^d	2.9	2.5	2.8	6.2	4.9	(3.9)
BES with significant hit	10,639	15,273	14,177	8,066	5,546	53,701

^a The number of BESs that have Q20 length is of 100 bp or greater. These figures are used for the evaluation of sequencing.

^b The number of BESs after multi-block setting was applied and bacterial sequences were removed.

^c The number of BESs after repetitive-rich sequences detected by *RepeatMasker* were removed.

^d The number of BESs *uniq-sequences* after *self-match* sequences were removed.

The “Q20 lengths” of the five species were very long (ranged from 588 bp to 614 bp). A BES with Q20 bases equal or longer than 100 bp was regarded as successful. The rate of the number of the successful sequences to the total number of BESs was called the “success rate.” The success rate is also a common indicator for quality assessment of sequencing (Zhao et al., 2001). The success rate for each species is shown in Table 1 (Success rate). They are as high as 88% to 93%, while the typical rates of other projects are 79% with the condition of high quality length > 200 bp (Larkin et al., 2003), and 81% with the condition of read length > 100 bp (Zhao et al., 2001). The average Q20 lengths ranged from 565 bp to 598 bp (Table 1, Average length). Total numbers of succeeded bases were shown in Table 1 (Total (bp)). A BES with Q20 length longer than 100 bp was regarded as a ‘successful’ sequence. The number of such BESs ranged from 14,341 to 20,355 (Table 1, Q20-HQ BESs). For most of the BAC libraries, both BESs are ‘successful’, although some BAC libraries have only one-side ‘successful’ BES (Table 1, BESs one-end, BESs mate-pairs). For *D. auraria* as an example, 977 BAC clones ‘succeeded’ for one-side BESs, and 7,304 BAC clones ‘succeeded’ for both BESs.

Repetitive Sequences After filtering vector sequences and bacterial sequences, a portion of the BESs were extracted in such a way that each sequence consists of continuous (at least 50 bp) high QV bases (QV ≥ 20). The

condition was fulfilled by the multi-block setting method (see Materials and Methods). As the high quality sequences, 77,165 BESs were extracted for all the species (Table 1, Q20-passed BESs). From the extracted high quality sequences, repetitive sequences were identified by the RepeatMasker program (version 2004/03/06 with RepBase 8.12) with options “-xsmall -species drosophila.” Detailed results are available at the WWW sites (<http://fly.gsc.riken.jp/>). The fraction of repetitive sequences ranged from 10% to 16% (Table 1, ‘Masked rate, b’). In the count, a BES that does not contain non-repetitive regions longer than 100 bp was designated *repeat*, because the BES was dominated by repeat sequences.

In addition, we performed intra-species self-comparison of the BESs with the NCBI-BLAST (sequence similarity search) program to identify putative repetitive sequences yet unknown and to remove them from the BES. A homologous region found by BLAST is called a *hit* here; this is called a high-scoring segment pair or HSP in BLAST theory. If a region in a BES has multiple hits with many different BESs in the self-comparison, they are very likely to be novel repetitive sequences or duplicated regions. However, since it is not easy to distinguish the former from the latter, we set a criterion to identify novel repetitive sequences, which are designated *multi-hits* (see Methods). We masked the *multi-hits* in the BESs data and selected BESs which contained continuous non-masked regions longer than 100 bp. Eventu-

ally, BESs ranging from 10,092 to 15,354 were used for further analysis for each species (Table 1, Passed BESs by MM) and designated *unique-sequences*. The BESs not selected at this stage are designated *self-matches*.

Comparative BES mapping and Coverage The *unique-sequences* were compared with the *D. melanogaster* release 3 genome sequences to identify which locations the BESs correspond to. BLASTN was performed with both the default parameter set and an arranged parameter set (see Methods). The BLAST with the arranged parameter set output 14 and 17% more significant matches than that with the 'default' parameter set for BESs of DAU1 and DAN1, respectively, which are the two distant species from the mapped species *D. melanogaster*. For two closely related species and *D. melanogaster* itself, there are no significant differences between the two parameter sets in the numbers of significant matches. As the result, the number of significant hits (at least 50 bp, e-value $1e-5$) ranged from 5,546 to 14,177 for each species (total 53,701, Table 1, BESs with significant hit). If the two BESs hits of a BAC clone are inward in direction and the distance L_{in} of them on the public *D. melanogaster* genome is reasonable ($L_{in} \leq 300$ kbp), the BESs were called *pair-end mapped* (PEM). A cluster of PEMs suggests that the region is a part of a synteny block or a long homologous region. A snapshot of the *pair-end mapped* BAC libraries is shown in Fig. 3. The statistics of the mapped BES are shown in Table 2 in accordance with chromosome arms of mapped DME1, and the corresponding Muller's elements. All the statistics concerning base pairs were calculated on the mapped *D. melanogaster* genome. Most of the region of the *D. melanogaster* genome was covered by the mapped BAC clones of the three species DME1 (94%), DSM1 (97%), and DSE1 (96%). On the other hand, the two phylogenetically distant species, DAU1 and DAN1, showed relatively low coverage (57% for DAU1, and 47% for DAN1); in particular,

no PEMs were found on chromosome 4. The low coverage rates by DAU1 and DAN1 are reflected in the high sequence divergence between these two species and *D. melanogaster* rather than the quality of the BAC libraries. Here we assume that there are no strong biases to affect the statistics in construction of BAC libraries, such as unbalanced distribution of a restriction enzyme site used in the construction of BAC libraries. The coverage of chromosome 4 of DME1 is less (69%) than other autosomes ($> 90\%$). This result could reflect that the chromosome 4 of *D. melanogaster* contains higher proportions of repetitive sequences than the other chromosomes. The chromosomes of *D. melanogaster* contain approximately seven times as much dispersed repetitive and nomadic DNA as those of the sibling species *D. simulans* (Dowsett and Young, 1982). The portion of the repetitive sequences detected by the RepeatMasker program was 27.3% for chromosome 4, while 5.6–9.4% for the others. The ratio of transposable elements in chromosome 4 is 4.2 times larger than that of the other chromosomes (Gonzalez et al., 2002). To take account of the repeat contents, we calculated the rate of total number of bases covered by the BAC inserts mapped in *D. melanogaster* genome against the number of bases of non-repetitive sequences in *D. melanogaster* genome for each chromosome arm. The rate (net coverage) should be around 1.0 or higher when the number of BACs is large enough to cover all the unique sequences. The rate for the chromosome 4 was 0.96 and the rates for the others ranged from 0.99 to 1.04. Thus the net coverage of chromosome 4 was at the same level as the others when we take account of the abundance of repetitive sequences. The total coverage of DSM1 and DSE1 were higher than that of DME1, because the clone numbers of BAC libraries of DSM1 and DSE1 were larger than the DME1 library. The rate of 'unmapped BAC clones' to 'the all BAC clones with *unique-sequences*' were 1.3% (DME1), 2.2% (DSM1) and 2.0% (DSE1), as shown in Table 3. More portions of BAC

Fig. 3. Mapped BESs on the Public *D. melanogaster* Genome Sequences. The first line represents the positions on the genome. The short vertical lines indicate the position of BESs. The break lines connecting the two vertical lines (paired BESs) indicate the inserts of the individual BAC clones. Mapped BAC clones meet the following conditions, 1) both BESs of a BAC clone are inward direction, 2) the distance of the BESs are less than 300 kb. This is the case of BAC libraries of DAU1. The Fig. was produced using the UCSC genome browser.

Table 2. Statistics of Mapped BESs, and estimated coverages

Chromosome arm ^a	Muller's element ^a	% coverage ^b	Fold coverage ^b	Mapped BAC libraries	Contigs	Average length ^b (bp)	Gap length (bp) ^b
DME1							
X	A	89.65	2.58	602	59	330,928	37,588
2L	B	95.92	3.81	906	27	789,316	32,371
2R	C	92.19	3.31	722	37	505,867	41,728
3L	D	96.34	3.76	954	27	833,261	30,506
3R	E	94.85	3.75	1,128	47	562,887	29,898
4	F	69.41	1.82	26	5	171,839	63,113
Total		93.65	3.45	4,338	202	541,417	35,650
DSM1							
X	A	95.50	4.06	855	24	866,640	39,226
2L	B	97.35	5.78	1,261	10	2,162,908	53,532
2R	C	96.46	5.61	1,067	15	1,305,550	44,969
3L	D	97.40	6.20	1,378	6	3,790,895	86,692
3R	E	99.05	6.14	1,639	11	2,511,413	22,104
4	F	86.57	3.32	41	2	535,805	55,420
Total		97.15	5.57	6,241	68	1,668,444	44,964
DSE1							
X	A	92.73	3.12	809	62	325,747	25,138
2L	B	96.79	5.06	1,259	19	1,131,826	35,662
2R	C	92.86	4.23	973	32	589,184	43,905
3L	D	98.18	4.80	1,271	16	1,432,902	25,046
3R	E	98.88	4.73	1,506	17	1,622,260	17,354
4	F	65.20	1.32	17	4	201,770	86,158
Total		95.79	4.38	5,835	150	745,779	31,505
DAU1							
X	A	32.26	0.56	147	62	113,315	234,198
2L	B	64.24	1.46	388	91	156,835	86,369
2R	C	57.45	1.35	315	68	171,533	125,196
3L	D	68.13	1.65	415	82	194,020	89,670
3R	E	59.03	1.51	450	100	164,633	113,144
Total		56.55	1.32	1,715	403	162,121	123,061
DAN1							
X	A	23.87	0.40	106	54	96,263	301,487
3R	B	49.17	0.97	239	89	122,744	125,486
3L	C	50.04	0.90	187	74	137,289	135,245
2R	D	55.73	1.16	283	76	171,229	134,271
2L	E	53.50	1.10	320	106	140,767	121,211
Total		46.92	0.92	1,135	399	135,881	151,800

^a The chromosome arms and the Muller's elements are those of each species.

^b These Figs. are estimated from the comparative BES (BAC) mapping results, when we assume that the lengths of the mapped BAC inserts of each species were the same as those of *D. melanogaster* and sequenced (The arm lengths of *D. melanogaster* are 21,780,003 bp for element A, 22,217,931 bp for B, 20,302,755 bp for C, 23,352,213 bp for D, 27,890,790 bp for E, and 1,237,870 bp for F, respectively.)

Table 3. Mapped and Unmapped BAC libraries

	DME1		DSM1		DSE1		DAU1		DAN1	
Pair-ends mapped	4,338	70.4%	6,241	70.8%	5,835	71.5%	1,715	31.6%	1,135	31.3%
Singleton	1,746	28.3%	2,380	27.0%	2,156	26.4%	2,607	47.7%	1,715	47.2%
Unmapped	81	1.3%	198	2.2%	166	2.0%	1,011	20.7%	781	21.5%
Total	6,165	100.0%	8,819	100.0%	8,157	100.0%	5,333	100.0%	3,631	100.0%

clones were mapped for DME1 than those of DSM1 and DSE1. When a paired BESs for a BAC clone significantly hits but the BESs are not PEM, the BAC clone was counted as ‘unmapped’.

Figure 4 shows the numbers of BESs categorized in the course of the analysis pipeline. The numbers of the *self-match* BESs (removed BESs at the step of self-comparison) of DAN1 are relatively larger than the others. This might be related to the fact that the number of the detected repetitive sequences in the BESs of DAN1 in Table 2 is smaller than for the other species. This implies that there are many unknown repetitive sequences in DAN1 that are not detected by the current RepeatMasker program.

We performed *in situ* hybridization using three BAC clones to identify what locations the BAC inserts came from in the species *D. ananassae*. The signals of the three BAC clones (DNB1-008L04, DNB1-016M14, and DNB1-001A14) were located at the cytogenetic regions of 33C, 52B, and 49D on salivary gland chromosomes, respectively, in the *D. ananassae* genome. On the other hand, these BAC clones were mapped onto the *D. melanogaster* genome *in silico* to the cytogenetic regions of 86D1–D3, 77D1–77E2, and 67F3–68A1. These results indicate that the homology of the chromosomal elements (Muller, 1940) between *D. melanogaster* and those of *D. ananassae* have been confirmed by *in situ* hybridization with the three BAC clones.

Fig. 4. Cumulative Numbers of BESs in Each Category for Each Species. The Y-axis represents the number of BESs. X-axis represents species of the BESs analyzed.

We calculated the ratio of the number of BESs that were mapped on exons to the total number of BESs that were mapped on the *D. melanogaster* genome. We used the annotation described in the UCSC genome browser (Kent et al., 2002) in which mRNA sequences in RefSeq database (Pruitt and Maglott, 2001) were mapped. The rates were between 12% and 16% for all the five species. This indicates that most of the mapped regions were non-exons. Thus we could perform comparative mapping successfully even for non-coding region for relatively distant species such as *D. melanogaster* and *D. ananassae*, which diverged approximately 44 million years ago (Tamura et al., 2004).

Analysis of BLAST hits and evolution We examined the significant BLASTN hits (HSP) of PEM-BESs to explore the evolution of the species' genome sequences. The features of the sequence alignment (percent identity, length, and negative logarithm of E-value) of BLAST hits give us different aspects of the alignments, although they are not independent of each other. The averages of the features are summarized in Table 4. The length of BLAST hits can be an indicator of homology, given that the average lengths of the BESs were almost the same among the species. The phylogenetic distances between *D. sechellia* and *D. melanogaster*, and between *D. simulans* and *D. melanogaster*, were not resolved well here. However, the BLAST analysis gives some clues. Even the lengths of BESs of DSM1 were slightly shorter than those of DSE1 on average, all the three features suggest that the mapped genomic sequences of *D. melanogaster* are more homologous to those of *D. simulans* than those of *D. sechellia* (Table 4). Statistically, Wilcoxon rank sum test showed p-values of 1.6×10^{-29} , 4.2×10^{-2} , and 1.9×10^{-4} ,

Table 4. Statistics of BLAST hits

	DME1	DSM1	DSE1	DAU1	DAN1
Identity	98.2	91.9	91.4	79.4	79.3
Length	520	465	460	265	245
Negative log ₁₀ e-value	188	160	157	52	46
BES length	521	506	508	510	494

The averages of the features of BLAST hits (percent identity, length, and negative log₁₀ e-value) and the average length of the BESs are shown from the comparative BES mapping. Only PEM-BESs were used in the analysis.

for percent identity, length, and e-value, respectively. This suggests that *D. sechellia* has undergone more divergence than *D. simulans* after the speciation of the ancestors. For a 57 kb regions (56F10–16), it was suggested that the noncoding sequences have been shortened by deletion in the genomes of *D. sechellia* and *D. simulans* (Kawahara et al., 2004). The number of the deletions were observed to be 7.5 times the number of insertions in *D. sechellia* (Kawahara et al., 2004). Therefore, the main reason of the divergence might be the deletions in the intergenic regions in *D. sechellia*.

The estimated insert size of the BAC libraries We have estimated the insert size of the BAC libraries as follows. Let L_{in} be the distance of two regions on the *D. melanogaster* genome that are the most homologous regions of the two BESs of a BAC clone. Here we consider only PEM BAC clones. Although the actual insert

size of the BAC libraries could be different from L_{in} to some extent, the L_{in} can be regarded as an estimate of the insert size of the BAC libraries. Most of the insert size ranged from 100 to 150 kb, which are typical lengths for the BAC inserts. For some portion of the BAC libraries, the insert size is below 50 kb. We measured the insert size by electrophoresis for some species. Since the insert sizes of 20% of the clones were actually below 50 kb, the BAC clone distribution obtained from *in silico* comparative BES mapping can be considered to be appropriate.

Sequences categorized as ‘self-matches’ are likely to contain novel repeats As shown in Fig. 4, the BESs categorized as ‘self-match’, which are intra-species sequence matches in self-comparison of BESs, in *D. ananassae* (DAN1) are relatively large. This may be related to the fact that chromosome 4 of DAN1 are almost exclusively composed of heterochromatin, and the heterochromatin part is much larger than those of other species. Besides, from ‘Masked rate’, Table 1, slightly small portion of the BESs of *D. ananassae* (DAN1, 10.7%) was identified as repetitive sequences by the RepeatMasker program. Therefore the BESs categorized as ‘self-match’ in *D. ananassae* might contain lots of novel repetitive

Fig. 5. Chromosomal Rearrangements and the Dot plot for Confirmation of Breakpoints. (a). The upper three tandem thick arrows (open, shaded, and solid) indicate the genomic sequences of species *S* (Genome *S*), from which BAC clones were constructed. Below them, the three thick arrows represent the reference genome sequences of species *T* (Genome *T* = *D. melanogaster*), which is used to map BESs. The shaded thick arrows indicate the syntenic regions at which an inversion occurred in either species. A pair of small arrows and a line indicate BAC end sequences and an insert, respectively. BESs with straight lines indicate valid BESs, while BESs with dotted lines indicate invalid BESs. When there is an inversion between *S* and *T*, there exist two breakpoints (vertical broken lines). Given the form of the BES-genome map, the existence and location of the inversion can be inferred. The two regions in X-axis (BP region.1 and region.2) were magnified in (b) and (c). The same contigs were mapped in the two regions, indicating that the two breakpoints were confirmed by the contigs. (b). The "dot plot" using the contigs of *D. simulans* (DSM1) and the genome sequences of *D. melanogaster* (DME1) corresponding to the predicted breakpoint region "BP region 1" (chromosome arm 3R, 3,856,802–3,896,705 in base positions; approximately 40 kb in length including 10 kb margins on both upstream and downstream). The Y-axes represent the base positions of alignments of the several contigs. The plot points indicate the endpoints of the BLAST alignments between contigs and the target genome sequences. The white arrow shows the breakpoint at around 18,000 bp in X-axis. (c). The "dot plot" using the contigs of *D. simulans* and the genome sequences of *D. melanogaster* corresponding to the predicted breakpoint region "BP region 2" (chromosome arm 3R, 17,546,012–17,585,421 in base positions; approximately 40 kb in length including 10 kb margins only on upstream). The white arrow shows the breakpoint at around 15,000 bp in X-axis.

sequences, which do not exist in the database of RepeatMasker. To explore the possibility, we extracted the sequences categorized as 'self-match', and clustered them by sequence similarity, and constructed a consensus sequence for each cluster. The consensus sequences were compared with the non-redundant nucleotide database, called 'nt' provided by NCBI (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>). Most of the consensus sequences matched best with un-annotated regions in *D. melanogaster* genome. They are likely to be novel repetitive regions. Parts of consensus sequences matched best with genes of different species of *Drosophila* with percent identity between 80% and 90%. This suggests that these BESs are parts of genes homologous to the matched genes. Most of the homologues may comprise a family with other paralogues in *Drosophila*, which are often found in multiple regions in the genome. Another possibility is that they are parts of pseudogenes. In fact, the two clusters of DAU1 matched with the same transposon mariner transposase pseudogene of *Bactrocera neohumeralis* (AF348438). Both matches are longer than 129 bp and similarity (percent identity) was higher than 86%.

Prediction of breakpoints of the synteny blocks from the comparative BES mapping By using the comparative BES mapping, we predicted breakpoints as boundaries of synteny blocks that were generated between one species and the *D. melanogaster* genome. When a pair of BESs of a BAC clone matched significantly to *D. melanogaster* genome, the BESs were categorized as either *valid* or *invalid* (See Materials and Methods, and Fig. 5). The *invalid* BESs imply that the BESs are involved in any chromosomal rearrangements. Using such BESs, we predicted the boundaries of genome alignments or syntenic regions, where some chromosomal rearrangements presumably have occurred. We predicted them with only BES-genome map, not using two genome sequences.

As the result, 111 estimated breakpoints (3, 2, 47, and 59 for DSM1, DSE1, DAU1, and DAN1, respectively) were found in the species examined (Appendix A, Table A). Between *D. melanogaster* and *D. simulans*, two major inversions are long known, one in chromosome 3R (Horton, 1939), and the other in chromosome 4 (Horton, 1939; Podemski et al., 2001). The two breakpoints found in this study are of the inversion in chromosome arm 3R. We did not detect any breakpoints in chromosome 4, because of few *invalid* BES hits in this chromosome. Genomic contigs of *D. simulans* sequence were taken from the WWW site of *D. simulans* genome sequencing project (http://www.genome.wustl.edu/genome_group.cgi?GROUP=6), and compared to *D. melanogaster* genome. There were long contigs of *D. simulans* covering the region in question. The dot plots of contigs from *D. simulans* against

two sequences from *D. melanogaster* are shown in Fig. 5. This Fig. indicates the presence of an inversion and distantly located two breakpoints that correspond well to the cytological breakpoints, 84E and 93E/F.

We thank Ehime University for providing *Drosophila* samples to construct BAC libraries. We also thank C. Kawagoe, T. Katayama, X. Son, and all technical staff of the Human Genome Research Group and the Genome Core Technology Facility, RIKEN Genomic Sciences Center for their excellent work and support. This work was supported in-part by a grant of the National Bio Resource Project from the Ministry of Education, Science, Sports, and Culture of Japan and a Special Fund for the RIKEN Genomic Sciences Center.

REFERENCES

- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.
- Altschul, S. F. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
- Bergman, C., Pfeiffer, B., Rincon-Limas, D., Hoskins, R., Gnirke, A., Mungall, C., Wang, A., Kronmiller, B., Pacleb, J., Park, S., et al. (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biology* **3**, research0086.0081–0086.0020.
- Bischof, J., Maeda, R. K., Hediger, M., Karch, F., and Basler, K. (2007) An optimized transgenesis system for *Drosophila* using germ-line-specific phiC31 integrases. *Proc. Natl. Acad. Sci. USA* **104**, 3312–3317.
- Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218.
- Dowsett, A. P., and Young, M. W. (1982) Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. *Proc. Natl. Acad. Sci. USA* **79**, 4570–4574.
- Ewing, B., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using *Phred*. II. Error Probabilities. *Genome Res.* **8**, 186–194.
- Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using *Phred*. I. Accuracy Assessment. *Genome Res.* **8**, 175–185.
- Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T. D., Itoh, T., Tsai, S. F., Park, H. S., Yaspo, M. L., Lehrach, H., Chen, Z., et al. (2002) Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295**, 131–134.
- Gilbert, D. G. (2007) DroSpeGe: rapid access database for new *Drosophila* species genomes. *Nucleic Acids Res* **35**, D480–D485.
- Gonzalez, J., Ranz, J. M., and Ruiz, A. (2002) Chromosomal Elements Evolve at Different Rates in the *Drosophila* Genome. *Genetics* **161**, 1137–1154.
- Higa, I., and Fuyama, Y. (1993) Genetics of food preference in *Drosophila sechellia*. I. Responses to food attractants. *Genetica* **88**, 129–136.
- Horton, I. H. (1939) A Comparison of the Salivary Gland Chromosomes of *Drosophila Melanogaster* and *D. Simulans*. *Genetics* **24**, 234–243.

- Kawahara, Y., Matsuo, T., Nozawa, M., Shin, I. T., Kohara, Y., and Aigaki, T. (2004) Comparative sequence analysis of a gene-dense region among closely related species of *Drosophila melanogaster*. *Genes Genet Syst* **79**, 351–359.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res* **12**, 996–1006.
- Kimura, M. T. (1988) Interspecific and geographic variation of diapause intensity and seasonal adaptation in the *Drosophila auraria* species complex (Diptera: Drosophilidae). *Functional Ecology* **2**, 177–183.
- Korf, I., Yandell, M., and Bendell, J. (2003) BLAST, Sebastopol, O'Reilly & Associates Inc.
- Larkin, D. M., Everts-van der Wind, A., Rebeiz, M., Schweitzer, P. A., Bachman, S., Green, C., Wright, C. L., Campos, E. J., Benson, L. D., Edwards, J., et al. (2003) A Cattle-Human Comparative Map Built with Cattle BAC-Ends and Human Genome Sequence. *Genome Res* **13**, 1966–1972.
- Leeb, T., Vogl, C., Zhu, B., de Jong, P. J., Binns, M. M., Chowdhary, B. P., Scharfe, M., Jarek, M., Nordsiek, G., Schrader, F., et al. (2006) A human-horse comparative map based on equine BAC end sequences. *Genomics* **87**, 772–776.
- Muller, H. J. (1940) Bearing of *Drosophila* work on systematics. In: *The New Systematics* (ed. J. Huxley), pp. 185–268. Oxford University Press (Clarendon), London.
- Oberstein, A., Pare, A., Kaplan, L., and Small, S. (2005) Site-specific transgenesis by Cre-mediated recombination in *Drosophila*. *Nat. Methods* **2**, 583–585.
- Olson, S. A. (2002) EMBOSSE opens up sequence analysis. *European Molecular Biology Open Software Suite. Brief Bioinform.* **3**, 87–91.
- Podemski, L., Ferrer, C., and Locke, J. (2001) Whole arm inversions of chromosome 4 in *Drosophila* species. *Chromosoma* **110**, 305–312.
- Pruitt, K. D., and Maglott, D. R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140.
- Ren, C., Lee, M.-K., Yan, B., Ding, K., Cox, B., Romanov, M. N., Price, J. A., Dodgson, J. B., and Zhang, H.-B. (2003) A BAC-Based Physical Map of the Chicken Genome. *Genome Res.* **13**, 2754–2758.
- Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M. J., Chen, R., Meisel, R. P., et al. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution. *Genome Res.* **15**, 1–18.
- Sawamura, K., Watanabe, T. K., and Yamamoto, M. T. (1993) Hybrid lethal systems in the *Drosophila melanogaster* species complex. *Genetica* **88**, 175–185.
- Sawamura, K., Karr, T. L., and Yamamoto, M. T. (2004) Genetics of hybrid inviability and sterility in *Drosophila*: dissection of introgression of *D. simulans* genes in *D. melanogaster* genome. *Genetica* **120**, 253–260.
- Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000) PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586.
- Tamura, K., Subramanian, S., and Kumar, S. (2004) Temporal Patterns of Fruit Fly (*Drosophila*) Evolution Revealed by Mutation Clocks. *Mol. Biol. Evol.* **21**, 36–44.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
- Tobari, Y. N. (1993) *Drosophila ananassae*. *Genetical and Biological Aspects.*, Tokyo and Basel, Switzerland, Japan Scientific Societies Press and Karger.
- Tomaru, M., Matsubayashi, H., and Oguma, Y. (1995) Heterospecific inter-pulse intervals of courtship song elicit female rejection in *Drosophila biauraria*. *Anim. Behav.* **50**, 905–914.
- Venken, K. J., He, Y., Hoskins, R. A., and Bellen, H. J. (2006) P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* **314**, 1747–1751.
- Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M., et al. (2002) Comparative Genome and Proteome Analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159.
- Zhao, S., Malek, J., Mahairas, G., Fu, L., Nierman, W., Venter, J. C., and Adams, M. D. (2000) Human BAC Ends Quality Assessment and Sequence Analyses. *Genomics* **63**, 321–332.
- Zhao, S., Shatsman, S., Ayodeji, B., Geer, K., Tsegaye, G., Krol, M., Gebregeorgis, E., Shvartsbeyn, A., Russell, D., Overton, L., et al. (2001) Mouse BAC Ends Quality Assessment and Sequence Analyses. *Genome Res.* **11**, 1736–1745.

Table A. Predicted Breakpoints

Species	Muller's element	Chromosome arms	Start positions	End positions	DAN1	B	chr3R	5,919,940	6,139,601
DSM1	C	chr2R	1,730,857	2,038,666	DAN1	B	chr3R	10,043,621	10,550,883
DSM1	E	chr3R	3,866,802	3,888,705	DAN1	B	chr3R	11,315,576	11,870,743
DSM1	E	chr3R	17,556,012	17,605,421	DAN1	B	chr3R	13,213,301	13,281,707
DSE1	E	chr3R	3,814,078	3,879,793	DAN1	B	chr3R	13,362,500	13,432,014
DSE1	E	chr3R	17,549,196	17,606,965	DAN1	B	chr3R	14,488,087	14,573,013
DAU1	B	chr2L	4,801,081	4,821,237	DAN1	B	chr3R	14,536,909	14,895,190
DAU1	B	chr2L	6,330,201	6,612,482	DAN1	B	chr3R	16,102,951	16,207,074
DAU1	B	chr2L	6,369,827	6,689,681	DAN1	B	chr3R	16,637,955	16,665,500
DAU1	B	chr2L	9,896,108	9,905,914	DAN1	B	chr3R	16,809,811	17,120,246
DAU1	B	chr2L	12,086,481	12,122,414	DAN1	B	chr3R	17,391,515	17,461,658
DAU1	B	chr2L	12,483,608	12,629,882	DAN1	B	chr3R	17,539,704	17,567,872
DAU1	B	chr2L	16,657,327	16,715,223	DAN1	B	chr3R	18,145,456	18,325,135
DAU1	B	chr2L	17,953,886	17,972,769	DAN1	B	chr3R	20,651,335	20,706,401
DAU1	B	chr2L	20,350,435	20,466,057	DAN1	B	chr3R	20,991,783	21,313,308
DAU1	B	chr2L	20,450,872	20,654,248	DAN1	B	chr3R	21,883,752	22,370,243
DAU1	B	chr2L	20,863,333	20,971,979	DAN1	B	chr3R	23,710,374	23,780,794
DAU1	C	chr2R	9,065,329	9,123,512	DAN1	B	chr3R	24,601,037	24,672,740
DAU1	C	chr2R	10,272,189	10,414,727	DAN1	B	chr3R	24,727,079	24,999,365
DAU1	C	chr2R	13,041,352	13,122,516	DAN1	C	chr3L	1,317,854	1,567,109
DAU1	C	chr2R	13,562,565	13,651,014	DAN1	C	chr3L	3,889,838	4,022,346
DAU1	C	chr2R	14,075,421	14,171,802	DAN1	C	chr3L	6,892,433	7,007,721
DAU1	C	chr2R	18,096,891	18,179,210	DAN1	C	chr3L	7,880,290	8,058,072
DAU1	D	chr3L	489,894	666,593	DAN1	C	chr3L	8,330,383	8,401,825
DAU1	D	chr3L	1,227,030	1,263,542	DAN1	C	chr3L	14,930,082	15,073,450
DAU1	D	chr3L	5,057,542	5,078,107	DAN1	C	chr3L	15,958,219	16,042,442
DAU1	D	chr3L	5,857,447	5,864,964	DAN1	C	chr3L	16,980,842	17,026,145
DAU1	D	chr3L	6,703,832	6,970,803	DAN1	C	chr3L	19,479,573	19,553,543
DAU1	D	chr3L	8,135,851	8,295,992	DAN1	C	chr3L	20,016,053	20,366,617
DAU1	D	chr3L	8,394,239	8,437,824	DAN1	D	chr2R	1,741,299	2,051,851
DAU1	D	chr3L	9,483,621	9,732,603	DAN1	D	chr2R	3,290,654	3,350,865
DAU1	D	chr3L	10,629,058	10,727,519	DAN1	D	chr2R	4,840,177	4,939,895
DAU1	D	chr3L	12,047,664	12,109,818	DAN1	D	chr2R	7,937,153	7,953,286
DAU1	D	chr3L	15,958,770	15,971,034	DAN1	D	chr2R	8,179,613	8,409,151
DAU1	D	chr3L	16,434,877	16,783,068	DAN1	D	chr2R	9,978,652	10,296,863
DAU1	D	chr3L	17,204,732	17,220,968	DAN1	D	chr2R	11,710,777	11,777,544
DAU1	E	chr3R	63,551	149,424	DAN1	D	chr2R	12,898,221	13,201,364
DAU1	E	chr3R	2,116,595	2,245,915	DAN1	D	chr2R	14,690,198	14,818,098
DAU1	E	chr3R	6,563,598	6,618,289	DAN1	D	chr2R	15,673,159	16,113,860
DAU1	E	chr3R	8,511,319	8,582,451	DAN1	D	chr2R	17,634,333	17,693,630
DAU1	E	chr3R	10,670,417	10,744,239	DAN1	D	chr2R	19,217,135	19,466,734
DAU1	E	chr3R	11,996,047	12,212,620	DAN1	E	chr2L	19,839,479	20,184,682
DAU1	E	chr3R	14,043,728	14,162,051	DAN1	E	chr2L	555,849	830,732
DAU1	E	chr3R	14,477,231	14,575,544	DAN1	E	chr2L	1,689,545	1,726,022
DAU1	E	chr3R	16,076,546	16,410,667	DAN1	E	chr2L	4,132,662	4,279,082
DAU1	E	chr3R	16,941,169	17,237,306	DAN1	E	chr2L	5,417,117	5,611,134
DAU1	E	chr3R	17,231,877	17,576,615	DAN1	E	chr2L	7,773,252	8,031,304
DAU1	E	chr3R	19,429,601	19,774,544	DAN1	E	chr2L	7,989,662	8,199,702
DAU1	E	chr3R	20,647,451	20,719,805	DAN1	E	chr2L	11,452,651	11,799,122
DAU1	E	chr3R	22,712,974	23,075,567	DAN1	E	chr2L	11,732,567	11,887,829
DAU1	E	chr3R	24,741,623	24,766,298	DAN1	E	chr2L	13,782,705	13,822,008
DAU1	E	chr3R	27,018,674	27,064,470	DAN1	E	chr2L	16,676,735	16,737,187
DAU1	E	chr3R	27,396,360	27,666,520	DAN1	E	chr2L	18,674,862	18,747,161
DAN1	B	chr3R	2,851,849	2,898,582	DAN1	E	chr2L	18,821,283	19,458,719
DAN1	B	chr3R	3,737,215	3,830,491					
DAN1	B	chr3R	4,384,226	4,422,676					
DAN1	B	chr3R	4,609,020	4,622,206					
DAN1	B	chr3R	4,786,707	4,856,021					

Each column represents species, chromosome, start position and end position. It is difficult to describe a breakpoint by specifying a point at the resolution of base pairs. It is more appropriate to describe a breakpoint by a region with start base position and end base position, where chromosomal rearrangement may have occurred. All the positional information is the base position on *D. melanogaster* genome sequence.