The Drosophila BAC resource

The 19 Genomes of Drosophila: a BAC Library Resource for Genus-wide and Genome Scale Comparative Evolutionary Research

Xiang Song§1, Jose Luis Goicoechea§1, Jetty S. S. Ammiraju§1, Meizhong Luo§1, Ruifeng He§, Jinke Lin§, So- Jeong Lee§, Nicholas Sisneros§, Tom Watts§§, David A. Kudrna§, Wolfgang Golser§, Elizabeth Ashley§, Kristi Collura§, Michele Braidotti§, Yeisoo Yu§, Luciano M. Matzkin§§,* Bryant F. McAllister**, Therese Ann Markow§§,*2, Rod A. Wing§§,*

§ Arizona Genomics Institute and BIO5 Institute, School of Plant Sciences, University of Arizona, 1657 E. Helen Street Tucson, AZ 85721 USA

§§ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721 USA

*Division of Biological Sciences, University of California, San Diego, La Jolla, CA, 92093, USA.

** Department of Biology, University of Iowa, Iowa City, IA 52242 USA

These authors contributed equally to this work

1 Corresponding authors

Copyright 2011.
The *Drosophila* BAC resource

Running Title: The *Drosophila* BAC resource

Key Words: *Drosophila*, BACs, Comparative Phylogenomics, Physical maps

Corresponding authors:

Therese Ann Markow,

Division of Biological Sciences, University of California, San Diego, La Jolla, CA, 92093, USA.

Phone: (858) 246-0095; Fax: (858) 534-7108; Email: tmarkow@ucsd.edu

Rod A. Wing,

Arizona Genomics Institute and BIO5 Institute, School of Plant Sciences, University of Arizona, 1657 E. Helen Street Tucson, AZ 85721 USA

Email: rwing@ag.arizona.edu Phone: 520-626-9595; Fax: 520-621-1259.
The Drosophila BAC resource

ABSTRACT

The genus Drosophila has been the subject of intense comparative phylogenomics characterization to provide insights into genome evolution under diverse biological and ecological contexts, and to functionally annotate the D. melanogaster genome, a model system for animal and insect genetics. Recent sequencing of 11 additional Drosophila species from various divergence points of the genus is a first step in this direction. However, to fully reap the benefits of this resource, the Drosophila community is faced with two critical needs: i.e. the expansion of genomic resources from a much broader range of phylogenetic diversity, and the development of additional resources to aid in finishing the existing draft genomes. To address these needs, we report the first synthesis of a comprehensive set of BAC resources for 19 Drosophila species from all three subgenera. Ten libraries were derived from the exact source used to generate 10 of the 12 draft genomes, while the rest were generated from a strategically selected set of species based on salient ecological and life history features, and their phylogenetic positions. The majority of the new species have at least one sequenced reference genome for immediate comparative benefit. This 19 BAC library set was rigorously characterized and shown to have: large insert sizes (125 to 168 kb), low non-recombinant clone content (0.3% to 5.3%), and deep coverage (9.1X - 42.9X). Further, we demonstrated the utility of this BAC resource for generating physical maps of targeted loci, refining draft sequence assemblies, and identifying potential genomic rearrangements across the phylogeny.
The Drosophila BAC resource

INTRODUCTION

The genus Drosophila contains approximately 2000 species of diverse morphology, ecology and behavior that are placed in three major lineages: subgenus Sophophora, subgenus Drosophila and subgenus Dorsilopa (Markow and O’Grady 2006, 2007). The most widely studied species in the genus, D. melanogaster, is firmly established as the premier model system for many biological research areas such as neurobiology, medicine and population biology (Rubin and Lewis 2000). Several other species in this genus, such as D. pseudoobscura and D. virillis, have also been utilized as genetic model systems particularly for evolutionary studies (Anderson et al. 1991; Popadic and Anderson 1994; Orr and Coyne 1989; Charlesworth et al. 1997; Vieira et al. 1997; Sweigart 2010).

Recently, the genomes of D. melanogaster and 11 other Drosophila species, whose most recent common ancestor occurred more than 45-50 million years ago, have been sequenced, assembled and annotated (Adams et al. 2000; Myers et al. 2000; Celniker et al. 2002; Richards et al. 2005; Drosophila 12 Genomes Consortium 2007; Gilbert 2007). Species were selected for genome sequencing partly based on their relationship with D. melanogaster. Nine of the twelve sequenced genomes were sampled from one subgenus, Sophophora, to which D. melanogaster belongs and the remaining three are from the Drosophila subgenus. These sequences have already greatly improved understanding of the evolution and regulation of eukaryotic genes and genomes through comparative analyses (Stark et al. 2007). However, to fully reap the benefits from this unique resource, the Drosophila community has faced with two critical needs: first, the development of additional genomics resources to aid in finishing the 11 existing draft genome sequences;
The *Drosophila* BAC resource

and second, the generation of additional genomic resources that encompass a much broader range of phylogenetic diversity.


Genome sequences are available for 10 of 19 species for which BAC libraries are constructed, some of which were instrumental in facilitating sequence assemblies (*Drosophila* 12 Genomes Consortium 2007), and they remain a high priority resource for improving and finishing several of the low coverage draft genome assemblies. BAC libraries for species without sequenced genomes present an important resource for positional cloning and large-scale targeted comparative genome analyses.

We selected 19 species within three lineages of the genus *Drosophila* for BAC library construction (Figure 1). These species shared a common ancestor approximately 40-60 million years ago (Powell 1997) and were selected because of their varied evolutionary distances from *D. melanogaster* and other sequenced species, their diverse ecologies and life history characters, and the fact that they can be reared in the laboratory and used in experimental work in the future. Ten BAC libraries were constructed as a resource for
The *Drosophila* BAC resource generating BAC end mate-pair sequence to assist in the assembly of whole-genome shotgun sequences, and for enabling future genomic research (*Drosophila* 12 Genomes Consortium 2007). Beyond those 10 species, we are interested in generating BAC library resources for representative species of lineages not yet targeted for sequencing but which fill in large phylogenetic gaps. The majority of these species have at least one previously sequenced reference genome for immediate comparative benefit. In addition, this new set of species facilitates the “ladder and constellation” approach of modified phylogenetic shadowing proposed by Clark *et al.* (http://flybase.org/static_pages/news/whitepapers/GenomesWP2003.pdf) for annotating genome data. In this approach ladder rungs constitute successively increasing divergence points and constellations are clusters of species attaching to these divergence points. This set of 19 BAC libraries documented here will further advance the genus *Drosophila* as an ideal eukaryotic comparative genomics system designed to: 1) provide sequencing resources for comparative annotation of the *D. melanogaster* genome; and 2) provide genomic resources for experimental investigation of gene function throughout the genus *Drosophila*.

**MATERIALS AND METHODS**

Fly culturing and embryo collection

Fly cultures were expanded on banana/opuntia medium (http://flyfood.arl.arizona.edu/opuntia.php3) and healthy sexually mature adult flies were introduced into plexiglass oviposition chambers kept on a 16:8 light/dark cycle at 24-
The *Drosophila* BAC resource

25°C with a relative humidity of 60-80%. Exceptions to this procedure were: *D. littoralis*, *D. novamexicana*, *D. americana*, *D. grimshawi* and *D. persimilis* cultures which were oviposited at 20-22°C, whereas *D. albomicans* was oviposited at 17°C. Medium for *D. sechellia* was supplemented with 0.5% (v/v) hexanoic acid and 0.5% (v/v) octanoic acid to stimulate oviposition. Oviposition medium for *D. grimshawi* was supplemented with 2% (w/v) methylparaben to prevent overgrowth of fungus. *Drosophila busckii* and *D. grimshawi* cultures were grown on Wheeler-Clayton medium ([http://flyfood.arl.arizona.edu/wheeler.php3](http://flyfood.arl.arizona.edu/wheeler.php3)). *Drosophila grimshawi* adults were separated by sex until day of placement in the oviposition chamber to enhance embryo production. Adult flies were allowed to oviposit on a given plate for as long as possible without larval hatch. This interval varied between four and 48 hours depending on the species. About 1.2-1.5 grams wet weight embryos were pooled in batches and stored at -80°C at the end of each oviposition session.

Nuclei preparation and BAC library construction

Embryos were gently homogenized in PBS buffer (0.76% NaCl, 4mM NaH$_2$PO$_4$, 9mM Na$_2$HPO$_4$, PH 7.0) using a Dounce Tissue Grinder (Wheaton Science), centrifuged at 4°C at 1,430g for 15 min and resuspended in PBS buffer. The suspension was then mixed with an equal volume of 1% InCert Agarose (CAMBREX, in PBS buffer) at 45°C and transferred into plug molds. Treatment of plugs to produce un-sheared megabase-size DNA was as described (Luo and Wing 2003). BAC libraries were constructed as previously described (Luo and Wing 2003; Ammiraju *et al.* 2006).
The *Drosophila* BAC resource

**BAC library characterization**

DNA from a random sample of 260-480 BAC clones from each library was isolated, restriction digested with *NotI*, and run on CHEF gels for insert size determination as previously described (Luo and Wing 2003; Ammiraju *et al.* 2006).

High colony density hybridization filters for each library were prepared using Genetix Q-bots (Genetix) as described previously (Luo et al. 2006; Ammiraju *et al.* 2006). Nine gene specific probes were chosen that represented all chromosomes of *D. melanogaster* (Tables S1 and S2). All probe DNA fragments were PCR amplified from the *D. mojavensis* genome and gel purified using a QIAEX II (Qiagen) kit. Table S1 lists the primer sequences used for each probe. Purified DNA fragments were sequenced and similarity searches were conducted to validate their specificity. Probes were prepared by labeling with $^{32}$P dCTP using a DecaprimeII random prime labeling kit (Ambion), and hybridizations were carried out as described by Ammiraju *et al.* (2006). Positive clones were picked, re-arrayed on to colony filters, followed by a secondary hybridization with individual probes.

**Fingerprinting and contig assembly**

Positive hybridization clones were fingerprinted using SNaPshot (Luo *et al.* 2003; Kim *et al.* 2008), and assembled into contigs with FPC v 8.5.2 (Soderlund *et al.* 2000; [www.agcol.arizona.edu](http://www.agcol.arizona.edu)) at a fixed tolerance value 4 and an initial Sulston score $1e^{-50}$ (Ammiraju *et al.* 2006)
The *Drosophila* BAC resource

BAC end sequencing and in silico analysis


All BESs were masked with Repeat Masker (version3.1.0) against a redundant repeat database with sequences obtained from fly base (www.FlyBase.org) and Repbase (www.girinst.org). These sequences were used to conduct BLAST analysis against the mitochondrial (NC_001709, 19517 bp) and nuclear genome sequences of *Drosophila melanogaster* (Build 5.1) and the freeze 1 genome assemblies from the remaining eleven species

The *Drosophila* BAC resource

*virilis* and *D. mojavensis* were used as pseudo-reference sequences for the *D. virilis* and *D. repleta* species group, respectively. BES from *D. albomicans* and *D. busckii* was compared to the *D. grimshawi* sequences.

In addition, similarity searches were conducted with complete gene sequences of each probe against the 12 *Drosophila* whole genome sequences (*Drosophila* 12 Genomes Consortium 2007). Homologs with a minimum alignment length of 100 bp and 75 % of nucleotide identity were retained for further analysis and for a comparison of their presence or absence in FPC derived contigs.

**RESULTS AND DISCUSSION**

*Drosophila* strain selection and genome sizes:

Several criteria were used for careful evaluation of the different *Drosophila* species strains used for BAC resource development in this study. First, all fly lines were inbred for a minimum of 8 generations by sib-sib mating to reduce the extent of heterozygosity and subsequently sequenced at six nuclear loci to verify homozygosity (data unpublished). Second, to minimize endosymbiont contamination (*Wolbachia* spp. and *Spiroplasma* spp.) at least 5 adult fly DNA samples from each species were screened with established protocols (Mateos *et al.* 2006). Finally, species identity was confirmed by both morphological and molecular approaches. When a suitable nuclear or mitochondrial DNA marker was known for a species, that marker was amplified, sequenced and validated. Additionally, salivary gland chromosomes from third instar larvae were prepared and inspected for inversion polymorphism microscopically. Only
homokaryotypic lines were used. All strains (Table 1) are deposited in the UC San Diego Drosophila Stock Center and are publicly available as a community resource.

Genome size of an organism is the most important factor in determining the depth of a genomic library (reviewed in Gregory 2005). Previously determined genome sizes (Bosco et al. 2007) were used in this study for estimating the coverage of the BAC libraries for different Drosophila species. Bosco et al. (2007) employed two nucleic-acid binding fluorescent dyes, propidium iodide (PI) and 4′,6-diamidino-2-phenylindole (DAPI), in conjunction with flow cytometry to determine genome sizes of 38 species of Drosophilidae, including the 12 sequenced Drosophila species (Drosophila 12 Genomes Consortium 2007).

The genome sizes of 15 of the 19 Drosophila species used in this study were based on the PI method and the remaining species (D. novamexicana, D. littoralis, D. repleta and D. busckii) genome sizes were based alone the DAPI method alone (for which the PI data was not available) (Table 1). Nine of the Drosophila species strains were not the same as the strains analyzed by Bosco et al. (2007). An important finding to consider, as reported by Bosco et al. (2007) and Gregory and Johnston (2008), is that DAPI may overestimate genome size which could affect the estimated genome coverage of these 4 libraries.

Genome sizes of two species, D. arizonae and D. albomicans, were not known, so the genome sizes of closest relatives D. mojavensis and D. immigrans, respectively were applied to estimate the tentative genome coverages of their respective BAC libraries. The genome sizes among the 19 Drosophila species varied by ~3.2 fold, with the smallest being D. mercatorum and the largest D. virilis (Table 1).
The *Drosophila* BAC resource

243

**BAC library construction and characterization:**

Three different restriction enzymes were used for BAC library construction: *Hind*III, *Bam*HI, and *Bst*YI. Fifteen of the 19 libraries were constructed from DNA partially digested with *Hind*III, followed by size selection and ligation into the *Hind*II site of pIndigoBAC536Swal (Ammiraju *et al.* 2006) (Table 1). Two libraries each were generated similarly from *Bam*HI (*D. ananassae* and *D. mojavensis*) and *Bst*YI (*D. virilis* and *D. americana*) restriction digests. All libraries, except for the *D. busckii* library (two ligations) were built from single ligations. The number of clones in the 19 BAC library set ranged between 11,520 to 55,296 (Table 1), which were arrayed into 384-well microtiter plates for long-term storage in -80°C freezers at the Arizona Genomics Institute’s (AGI) BAC/EST Resource Center [www.genome.arizona.edu](http://www.genome.arizona.edu).

Insert sizes of individual clones in each library ranged from 10 kb to 371 kb, with the majority over 120 kb (Figure 2). The average insert sizes of these libraries ranged from 125 to 168 kb (Table 1). Percentages of non-insert containing clones ranged between 0.3% - 5.3%, which is typical for BAC libraries constructed at AGI (Ammiraju *et al.* 2006).

Genomic redundancy of the *Drosophila* BAC libraries

We estimated the genomic depth of the 19 *Drosophila* BAC set by three different, but complementary approaches. First, we estimated the redundancy of each library empirically from the average insert size, total number of clones, and the genome size of the corresponding lineage, which ranged approximately between 5.7 - 32.8 fold (Table 1).
To assess the randomness and extent of representational heterogeneity for different genomic regions, we screened the entire set of 19 *Drosophila* BAC libraries with 9 gene specific probes in two successive rounds of hybridizations (methods; Tables S1 and S2).

In brief, 4196 putative positive BAC clones were identified in the first round of hybridization, 3809 (91%) were confirmed by a second hybridization. The number of positive hits per library ranged from 1 to 108 (Table S3). At least one positive hit per each probe was detected for all the libraries with the exceptions of the *D. americana*, *D. repleta*, *D. hydei* libraries for probe X-CG11387 and *D. ananassae* for probe 3R-CG31247 (Table S3). In these four species no hits were found, even upon three rounds of library screening, with different hybridization stringencies. For *D. ananassae*, the whole genome draft sequence was available [http://rana.lbl.gov/Drosophila/caf1.html](http://rana.lbl.gov/Drosophila/caf1.html), and similarity searches revealed the presence of the probe sequence (3R-CG31247; Table S2) in the draft sequence assembly. Therefore, at least in the case of *D. ananassae*, it appears that methodological and/or library coverage issues prevented recovery of this gene via the hybridization based approach, possibly due to use of heterologous probes, multiple usage of high density colony filters, or cloning bias (under and over representation of genomic regions due to usage of a single restriction enzyme during library construction). More data is required to confirm the absence of the gene X-CG11387 in other three species (*D. americana*, *D. repleta*, *D. hydei*).

Hybridization based genome coverage’s ranged from 9.1X (*D. americana*) to 42.9X (*D. hydei*). In only two species, *D. mercatorum* and *D. willistoni*, the hybridization based coverage was slightly lower than expected (Table 2). The remaining 17 libraries either had nearly equal or higher coverage than predicted (Table 2, Table S3). The *D.
The *Drosophila* BAC resource

*a* albomicans BAC library showed a ~3.6 fold higher than expected coverage based on hybridization (Table 2), which could have resulted from not having accurate genome size estimation for this species (Table 1).

A third and a more rigorous approach using fingerprinted contig (FPC) based estimations of genomic redundancy of BAC libraries was applied using a similar strategy as our previous analysis of a set of 11 *Oryza* (cultivated and wild rice) BAC libraries (Ammiraju et al. 2006). This approach can discriminate the unavoidable cloning bias from those of cross hybridizations and genetic rearrangements such as duplications. All 3809 hybridization derived BAC clones were fingerprinted and 3005 (79%) successful fingerprints were assembled into physical contigs (Tables S4 and S5). Under a scenario of single copy probes and one contig per probe for each species, the theoretically expected number of contigs is 171 (9 probes for 19 libraries). However several exceptions were found; a) as described above, 1 probe X-CG11387 had no hits in the *D. americana; D. repleta* and *D. hydei* libraries, and another - 3R-CG31247 - had no hits in the *D. ananassae* library (Table S3); b) clones detected from 6 hybridizations (*D. yakuba, D. persimilis* and *D. willistoni* with probe X-CG11387; *D. mercatorum* with probe 2L-CG4128; *D. mercatorum, D. grimshawi* with probe 4-CG2999) resulted in the presence of singletons (Table S5) (all these instances resulted in less than 3 positive clones, Table S3). Taking into account the absence of these contigs in these species, 161 contigs are expected.

Our FPC analysis revealed a total of 211 contigs, 50 additional contigs than the expected number of 161 (Table S4). The number of contigs and respective coverage differed among different *Drosophila* libraries for the same probe (Table S5). Five probes
The *Drosophila* BAC resource (X-CG11387, X-CG32611, 3L-CG10948, 3R-CG31247, 4-CG2999) essentially behaved as single copy probes in most *Drosophila* libraries (Table S5). The remaining four detected on average, 1.4 or more contigs/per probe (Table S5). To better understand if these deviations from expectation (50 additional contigs) were due to technical issues (cross hybridization and assembly artifacts,) and/or lineage specific genetic changes, we gathered data from two additional experiments. First, based on BES mapping information (methods), we classified 142 contigs as primary (those that map to the expected genomic location) and 69 additional contigs as secondary (27 contigs that cannot be positioned in any genome and 42 contigs that map to non-orthologous locations), a good agreement between the results of FPC analyses and mapping information (Tables S2 and S6).

Second, nucleotide and protein similarity searches of the probe (or gene) sequences revealed that several secondary sites (17/42 secondary contigs) contained small cross hybridizing paralogous sequences (Table S6, indicated with *). It is possible that the 25 remaining secondary sites also contained very small cross hybridizing sequences that were not easily detected through similarity searches. In addition, sequence analysis of the extended flanking sequences of the primary sites with the secondary sites revealed no evidence of synteny, suggesting cross hybridization as the main cause for these additional contigs.

To provide a conservative estimate of genome coverage, we considered each identified contig as an independent locus and calculated a weighted FPC coverage that accounts for the presence of several loci (Table S4; Ammiraju *et al.* 2006). Estimated FPC coverage
The *Drosophila* BAC resource

for the 19 libraries (Table 2 and Table S4), ranged between 7 to 37X. Only two libraries had coverage below nine fold: *D. willistoni* (7X) and *D. americana* (8X).

Twelve libraries showed a ratio close to 1:1 between the FPC and empirically estimated coverage (Table 2). The *D. willistoni, D. littoralis, D. repleta, D. mercatorum, D. mojavensis, D. arizonae* and *D. busckii* libraries showed ratios equal or below 0.7:1 (Table 2; Table S4). The difference between hybridization based and contig based estimates of library coverage is due to the difference in the number of loci used to calculate the coverage. While each probe is considered as a single locus in the hybridization based approach, each secondary contig is considered as an independent locus in the FPC based approach (Table 2; Tables S3 and S4). Together, these results showcase the high quality and deep representational coverage of each of 19 *Drosophila* genomes in their respective libraries.

**Utilization of BAC libraries**

Although a few *Drosophila* BAC libraries have already been reported in the literature (Hoskins *et al.* 2000; Locke *et al.* 2000; Gonzalez *et al.* 2005; Osoegawa *et al.* 2007; Murakami *et al.* 2008), this is the first synthesis and characterization of a comprehensive set of BAC library resources for the genus, which fills a critical void for the *Drosophila* research community. Hybridization of nine different probes to the full set of libraries demonstrates the feasibility of isolating homologous regions across the entire genus. Combined with high-throughput sequencing methods (Wicker *et al.* 2006), this set of libraries provides an excellent resource for comparative studies of targeted genomic regions (e.g., Leung *et al.* 2010).
First, BAC libraries from species that do not yet have a reference genome sequence themselves provide a source for identifying genome rearrangements in comparisons with the available genome sequences. For example, end sequences of BACs isolated with the X-linked probe CG32611 from \textit{D. novamexicana} map at an unexpected position within contig 12970 of \textit{D. virilis}, indicating a putative small inversion at the base of the X chromosome that had not been previously identified (Vieira \textit{et al.} 1997). Another putative inversion was also revealed in \textit{D. arizonae} by the localization of end sequences of clones hybridizing to CG3139 in the genome sequence of \textit{D. mojavensis}. Targeted analyses inversion breakpoints are also enabled by the availability of these BAC libraries and informed by the reference genome sequences. Evans \textit{et al.} (2007) used cytological evidence on the position of an inversion in \textit{D. americana} to develop probes for isolating its breakpoints from the respective BAC clones. In addition, the BAC libraries for the nine un-sequenced \textit{Drosophila} species provide robust templates for the whole genome physical and sequence frameworks. In this direction, the entire \textit{D. persimilis} BAC library was fingerprinted, bidirectionally end sequenced, and assembled into a whole genome physical map. This map was aligned to the \textit{D. persimilis} and \textit{D. pseudoobscura} draft sequences, and is currently under editing (data not shown).

An extremely important application of the BAC resources reported here is in the ability to use functional genomics to test genes underlying the differences between \textit{Drosophila} species. The tool kit for functional analyses of \textit{Drosophila} has taken a major leap forward with the recent establishment of the P/\textsc{phi}C31 artificial chromosome manipulation (P[acman]) transgenesis platform (Venken \textit{et al.} 2006, 2007, 2009). While still reliant on the P transposable element for transformation, this BAC transgenic system
The *Drosophila* BAC resource

significantly improves upon the size of the DNA to be carried in the vector (>130 Kb), and it’s site specific integration in the fly genome. An important feature of the P[acman] system is recombinerring – which permits cloning/transfer of large DNA fragments from existing *Drosophila* P1 or BAC clones through a homologous recombination mediated gap repair process. Therefore, a combination of the P[acman] system with the 19 *Drosophila* BAC libraries will provide an unprecedented opportunity to the fly community to access, transfer and manipulate virtually any genomic region of interest (large genes or even gene clusters) covering the entire phylogenomic range of the genus *Drosophila*.

Finally, the BAC library set reported here can be used to further improve many of the existing *Drosophila* draft sequence assemblies (*Drosophila* 12 Genomes Consortium 2007), and aid in the characterization of lineage specific rearrangements. For example, physical mapping of BAC contigs, or individual BAC clones, identified by hybridization probes designed from draft *Drosophila* genome sequences, has revealed and confirmed chromosomal location of several sequence contigs from the draft assemblies, as well as their relationship to *D. melanogaster* (Table S6). Conserved linkage and physical markers were used to infer the physical organization of the assembled genome assemblies relative to reference chromosome maps (Schaeffer *et al.* 2008), and these BAC libraries serve as an appropriate resource to isolate regions at inferred gaps between adjacent contigs (e.g., Hoskins *et al.* 2000). Using hybridization to recover genome regions containing target genes, combined with end sequencing of positive clones further reveals the conserved linkage among *Drosophila* species. For example, scaffolds 20 and 24 map to X[A], 29 to 3L[D] and 30 to 4[F] in *D. sechellia*, 4512 4[F] in *D. erecta*, 12984 3R[B] and 12947
The *Drosophila* BAC resource


These libraries are likely to facilitate a wide array of comparative, evolutionary and functional genomics studies and play a major role in advancing the *Drosophila* biology.

409

ACKNOWLEDGEMENTS

410 This work was supported by the National Institute of Health grant #: U1HG02525A

411
The *Drosophila* BAC resource

**LITERATURE CITED**


AMMIRAJU, J. S., M. LUO, J. L. GOICOECHEA, W. WANG, D. KUDRNA *et al.*, 2006 The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. Genome Res. 16: 140–147.


The *Drosophila* BAC resource


The *Drosophila* BAC resource


The *Drosophila* BAC resource


KIM, H., B. HURWITZ, Y. YU, K. COLLURA, M. GILL *et al.*, 2008 Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*. Genome Biol. 9: R45.


The *Drosophila* BAC resource

A physical map of the polytenized region (101EF-102F) of chromosome 4 in *Drosophila melanogaster*. Genetics 155: 1175-1183.


The *Drosophila* BAC resource


ORR, H. A., and J. A. COYNE, 1989 The genetics of postzygotic isolation in the
The *Drosophila* BAC resource

*Drosophila virilis* group. Genetics 121: 527-537.


The *Drosophila* BAC resource


THE INTERNATIONAL HUMAN GENOME MAPPING CONSORTIUM, 2000a A physical map of the human genome. Nature 409: 934-941


The *Drosophila* BAC resource


WICKER, T., E. SCHLAGENHAUF, A. GRANER, T. J. CLOSE, B. KELLER et al., 2006 454 sequencing put to the test using the complex genome of barley. BMC genomics 7:275.
The *Drosophila* BAC resource

**FIGURE 1.** Phylogenetic tree of 19 species and *D. melanogaster* selected for the *Drosophila* BAC resource project. The phylogenetic relationships and approximate divergence times among the *Drosophila* species in our study were determined from a compilation of prior analyses (Pitnick et al. 1995; Markow and O’Grady 2006; *Drosophila* 12 Genomes Consortium 2007).

**FIGURE 2.** Insert size distribution of 19 *Drosophila* BAC libraries. Histograms A to S depict the insert size distribution in the 19 different libraries. For each histogram, X axis represents insert size (kb) and Y axis represents the number of clones in a particular insert size range.

A: *D. simulans* (DS_ABa); Average Insert Size 158 kb; B: *D. sechellia* (DS_Ba); Average Insert Size 139 kb; C: *D. yakuba* (DY_Ba); Average Insert Size 148 kb; D: *D. erecta* (DE_TBa); Average Insert Size 149 kb; E: *D. ananassae* (DA_Ba); Average Insert Size 148 kb; F: *D. persimilis* (DP_Ba); Average Insert Size 151 kb; G: *D. willistoni* (DW_Ba); Average Insert Size 150 kb; H: *D. americana* (DA_ABa); Average Insert Size 136 kb; I: *D. novamexicana* (DN_Ba); Average Insert Size 155 kb; J: *D. virilis* (DV_VBa); Average Insert Size 127 kb; K: *D. littoralis* (DL_Ba); Average Insert Size 168 kb; L: *D. repleta* (DR_Ba); Average Insert Size 143 kb; M: *D. mercatorum* (DM_Ba); Average Insert Size 125 kb; N: *D. mojavensis* (DM_CBa); Average Insert Size 143 kb; O: *D. arizonea* (DA_CBa); Average Insert Size 133 kb; P: *D. hydei* (DH_Ba); Average Insert Size 146 kb; Q: *D. grimshawi* (DG_Ba); Average Insert Size 127 kb; R: *D. albomicans* (DA_BBa); Average Insert Size 130 kb; S: *D. busckii* (DB_Ba); Average Insert Size 166 kb.
The *Drosophila* BAC resource

<table>
<thead>
<tr>
<th>Species</th>
<th>Group</th>
<th>Stock number</th>
<th>Library name</th>
<th>Enzyme</th>
<th>Genome size (Mb)</th>
<th>Average insert size (Kb)</th>
<th>Clone number</th>
<th>Calculated genome coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>D. simulans</em></td>
<td>MEL</td>
<td>DSSC# 14021-0251.195</td>
<td>DS_ABa</td>
<td>HindIII</td>
<td>160&lt;sup&gt;a&lt;/sup&gt;</td>
<td>158</td>
<td>18432</td>
<td>18.2</td>
</tr>
<tr>
<td><em>D. sechellia</em></td>
<td>MEL</td>
<td>DSSC # 14021-0248.25</td>
<td>DS__Ba</td>
<td>HindIII</td>
<td>166&lt;sup&gt;a&lt;/sup&gt;</td>
<td>139</td>
<td>18432</td>
<td>15.4</td>
</tr>
<tr>
<td><em>D. yakuba</em></td>
<td>MEL</td>
<td>DSSC# 14021-0261.01</td>
<td>DY__Ba</td>
<td>HindIII</td>
<td>188&lt;sup&gt;a&lt;/sup&gt;</td>
<td>148</td>
<td>11520</td>
<td>9.1</td>
</tr>
<tr>
<td><em>D. erecta</em></td>
<td>MEL</td>
<td>DSSC #14021-0224.01</td>
<td>DE_TBa</td>
<td>HindIII</td>
<td>145&lt;sup&gt;a&lt;/sup&gt;</td>
<td>149</td>
<td>18432</td>
<td>18.9</td>
</tr>
<tr>
<td><em>D. ananassae</em></td>
<td>MEL</td>
<td>DSSC # 14024-0371.13</td>
<td>DA_Ba</td>
<td>BanH</td>
<td>215&lt;sup&gt;a&lt;/sup&gt;</td>
<td>148</td>
<td>36864</td>
<td>25.4</td>
</tr>
<tr>
<td><em>D. persimilis</em></td>
<td>OBS</td>
<td>DSSC# 14011-0111.49</td>
<td>DP__Ba</td>
<td>HindIII</td>
<td>183&lt;sup&gt;a&lt;/sup&gt;</td>
<td>151</td>
<td>18432</td>
<td>15.2</td>
</tr>
<tr>
<td><em>D. willistoni</em></td>
<td>WIL</td>
<td>DSSC# 14030-0811.24</td>
<td>DW__Ba</td>
<td>HindIII</td>
<td>206&lt;sup&gt;a&lt;/sup&gt;</td>
<td>150</td>
<td>18432</td>
<td>13.4</td>
</tr>
<tr>
<td><em>D. americana</em></td>
<td>VIR</td>
<td>DSSC #15010-0951.15</td>
<td>DA_ABa</td>
<td>BstYI</td>
<td>275&lt;sup&gt;a&lt;/sup&gt;</td>
<td>136</td>
<td>11520</td>
<td>5.7</td>
</tr>
<tr>
<td><em>D. novamexicana</em></td>
<td>VIR</td>
<td>DSSC # 15010-1031.14</td>
<td>DN_Ba</td>
<td>HindIII</td>
<td>244&lt;sup&gt;b&lt;/sup&gt;</td>
<td>155</td>
<td>13440</td>
<td>8.5</td>
</tr>
<tr>
<td><em>D. virilis</em></td>
<td>VIR</td>
<td>DSSC # 15010-1051.87</td>
<td>DV_VBa</td>
<td>BstYI</td>
<td>404&lt;sup&gt;a&lt;/sup&gt;</td>
<td>127</td>
<td>55296</td>
<td>17.4</td>
</tr>
<tr>
<td><em>D. littoralis</em></td>
<td>VIR</td>
<td>DSSC # 15010-1001.11</td>
<td>DL__Ba</td>
<td>HindIII</td>
<td>238&lt;sup&gt;b&lt;/sup&gt;</td>
<td>168</td>
<td>36864</td>
<td>26</td>
</tr>
<tr>
<td><em>D. repleta</em></td>
<td>REP</td>
<td>DSSC# 15084-1611.10</td>
<td>DR__Ba</td>
<td>HindIII</td>
<td>167&lt;sup&gt;b&lt;/sup&gt;</td>
<td>143</td>
<td>36864</td>
<td>31.6</td>
</tr>
<tr>
<td><em>D. mercatorum</em></td>
<td>REP</td>
<td>DSSC# 15082-1521.36</td>
<td>DM__Ba</td>
<td>HindIII</td>
<td>128&lt;sup&gt;a&lt;/sup&gt;</td>
<td>125</td>
<td>18432</td>
<td>18</td>
</tr>
<tr>
<td><em>D. mojavensis</em></td>
<td>REP</td>
<td>DSSC # 15081-1352.22</td>
<td>DM_CBa</td>
<td>BanH</td>
<td>152&lt;sup&gt;a&lt;/sup&gt;</td>
<td>143</td>
<td>30720</td>
<td>28.9</td>
</tr>
<tr>
<td><em>D. arizonae</em></td>
<td>REP</td>
<td>DSSC# 15081-1271.27</td>
<td>DA_CBa</td>
<td>HindIII</td>
<td>152&lt;sup&gt;c&lt;/sup&gt;</td>
<td>133</td>
<td>18432</td>
<td>16.1</td>
</tr>
<tr>
<td><em>D. hydei</em></td>
<td>REP</td>
<td>DSSC# 15085-1641.58</td>
<td>DH_Ba</td>
<td>HindIII</td>
<td>164&lt;sup&gt;a&lt;/sup&gt;</td>
<td>146</td>
<td>36864</td>
<td>32.8</td>
</tr>
<tr>
<td><em>D. grimshawi</em></td>
<td>HAW</td>
<td>DSSC # 15287-2541.00</td>
<td>DG__Ba</td>
<td>HindIII</td>
<td>231&lt;sup&gt;a&lt;/sup&gt;</td>
<td>127</td>
<td>18432</td>
<td>10.1</td>
</tr>
<tr>
<td><em>D. albomicans</em></td>
<td>IMM</td>
<td>DSSC # 15112-1751.08</td>
<td>DA_BBa</td>
<td>HindIII</td>
<td>299&lt;sup&gt;c&lt;/sup&gt;</td>
<td>130</td>
<td>18432</td>
<td>8</td>
</tr>
<tr>
<td><em>D. busckii</em></td>
<td>DOR</td>
<td>DSSC# 13000-0081.31</td>
<td>DB_Ba</td>
<td>HindIII</td>
<td>194&lt;sup&gt;b&lt;/sup&gt;</td>
<td>166</td>
<td>18432</td>
<td>15.8</td>
</tr>
</tbody>
</table>

<sup>a</sup>Genome size measured by PI method (Bosco et al. 2007)

<sup>b</sup>Genome size measured by DAPI method a (Bosco et al. 2007)

<sup>c</sup>Genome sizes of *D. arizonae* and *D. albomicans* were adopted from the genome size of a close relatives, *D. mojavensis* and *D. immigrans*, respectively.

<sup>d</sup>Calculated genome coverage: by insert size, genome size and no of clones in the library

<sup>e</sup>MEL: melanogaster; OBS: obscura; WIL: willistoni; VIR: virilis; REP: repleta; HAW: Hawaiian; IMM: immigrans; DOR: subgenus *Dorsilopha*

<sup>f</sup>DSSC: *Drosophila* Species Stock Center
The *Drosophila* BAC resource

**TABLE 2** A comparison of genomic redundancies of each *Drosophila* BAC library as estimated by empirical, hybridization and by FPC approaches.

<table>
<thead>
<tr>
<th>Species</th>
<th>Calculated Genome Coverage</th>
<th>Average Hyb Coverage</th>
<th>FPC-General</th>
<th>Ratio of a:b:c</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>D. simulans</em></td>
<td>18.2</td>
<td>25.0</td>
<td>17</td>
<td>1 : 1.4 : 0.94</td>
</tr>
<tr>
<td><em>D. sechellia</em></td>
<td>15.4</td>
<td>20.2</td>
<td>14</td>
<td>1 : 1.3 : 0.88</td>
</tr>
<tr>
<td><em>D. yakuba</em></td>
<td>9.1</td>
<td>11.0</td>
<td>9</td>
<td>1 : 1.2 : 1.01</td>
</tr>
<tr>
<td><em>D. erecta</em></td>
<td>18.9</td>
<td>19.7</td>
<td>14</td>
<td>1 : 1.0 : 0.75</td>
</tr>
<tr>
<td><em>D. ananassae</em></td>
<td>25.4</td>
<td>25.3</td>
<td>22</td>
<td>1 : 1.0 : 0.87</td>
</tr>
<tr>
<td><em>D. persimilis</em></td>
<td>15.2</td>
<td>18.3</td>
<td>13</td>
<td>1 : 1.2 : 0.86</td>
</tr>
<tr>
<td><em>D. willistoni</em></td>
<td>13.4</td>
<td>9.6</td>
<td>7</td>
<td>1 : 0.7 : 0.52</td>
</tr>
<tr>
<td><em>D. americana</em></td>
<td>5.7</td>
<td>9.1</td>
<td>8</td>
<td>1 : 1.6 : 1.36</td>
</tr>
<tr>
<td><em>D. novamexicana</em></td>
<td>8.5</td>
<td>14.8</td>
<td>13</td>
<td>1 : 1.7 : 1.48</td>
</tr>
<tr>
<td><em>D. virilis</em></td>
<td>17.4</td>
<td>32.7</td>
<td>19</td>
<td>1 : 1.9 : 1.11</td>
</tr>
<tr>
<td><em>D. littoralis</em></td>
<td>26</td>
<td>25.1</td>
<td>18</td>
<td>1 : 1.0 : 0.71</td>
</tr>
<tr>
<td><em>D. repleta</em></td>
<td>31.6</td>
<td>35.7</td>
<td>14</td>
<td>1 : 1.1 : 0.44</td>
</tr>
<tr>
<td><em>D. mercatorum</em></td>
<td>18</td>
<td>11.7</td>
<td>10</td>
<td>1 : 0.6 : 0.54</td>
</tr>
<tr>
<td><em>D. mojavensis</em></td>
<td>28.9</td>
<td>31.1</td>
<td>17</td>
<td>1 : 1.1 : 0.59</td>
</tr>
<tr>
<td><em>D. arizonae</em></td>
<td>16.1</td>
<td>20.2</td>
<td>10</td>
<td>1 : 1.3 : 0.63</td>
</tr>
<tr>
<td><em>D. hydei</em></td>
<td>32.8</td>
<td>42.9</td>
<td>37</td>
<td>1 : 1.3 : 1.12</td>
</tr>
<tr>
<td><em>D. grimshawi</em></td>
<td>10.1</td>
<td>14.2</td>
<td>9</td>
<td>1 : 1.4 : 0.87</td>
</tr>
<tr>
<td><em>D. albomicans</em></td>
<td>8</td>
<td>28.4</td>
<td>10</td>
<td>1 : 3.6 : 1.22</td>
</tr>
<tr>
<td><em>D. busckii</em></td>
<td>15.8</td>
<td>28.2</td>
<td>9</td>
<td>1 : 1.8 : 0.58</td>
</tr>
</tbody>
</table>

*Theoretical coverage of each *Drosophila* library from the Table1*

*Average hybridization coverage; total number of clones detected by two rounds of hybridization divided by the total number of loci; from Table S3*

*FPC based estimate of genomic redundancy of each *Drosophila* library. Total number clones in each FPC assembly divided by the total number of contigs; from Tables S4 and S5.*