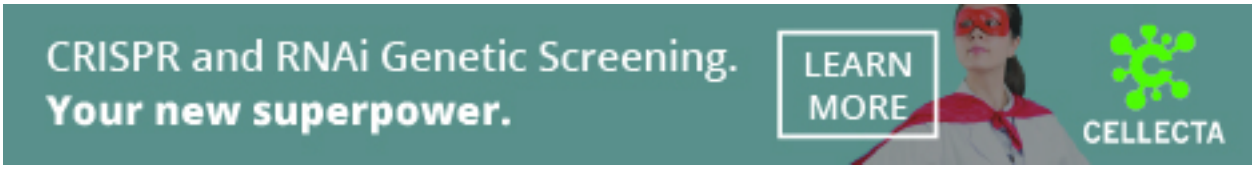
# Uprobe: A genome-wide universal probe resource for comparative physical mapping in vertebrates

Wendy A. Kellner, Robert T. Sullivan, Brian H. Carlson, et al.

*Genome Res.* 2005 15: 166-173
Access the most recent version at doi:10.1101/gr.3066805





# Uprobe: A genome-wide universal probe resource for comparative physical mapping in vertebrates

Wendy A. Kellner,[1] Robert T. Sullivan,[1] Brian H. Carlson,[1]
NISC Comparative Sequencing Program,[2] and James W. Thomas[1,3]

*[1]Emory University School of Medicine, Department of Human Genetics, Atlanta, Georgia 30322, USA; [2]Genome Technology Branch and NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*

Interspecies comparisons are important for deciphering the functional content and evolution of genomes. The expansive array of >70 public vertebrate genomic bacterial artificial chromosome (BAC) libraries can provide a means of comparative mapping, sequencing, and functional analysis of targeted chromosomal segments that is independent and complementary to whole-genome sequencing. However, at the present time, no complementary resource exists for the efficient targeted physical mapping of the majority of these BAC libraries. Universal overgo-hybridization probes, designed from regions of sequenced genomes that are highly conserved between species, have been demonstrated to be an effective resource for the isolation of orthologous regions from multiple BAC libraries in parallel. Here we report the application of the universal probe design principal across entire genomes, and the subsequent creation of a complementary probe resource, Uprobe, for screening vertebrate BAC libraries. Uprobe currently consists of whole-genome sets of universal overgo-hybridization probes designed for screening mammalian or avian/reptilian libraries. Retrospective analysis, experimental validation of the probe design process on a panel of representative BAC libraries, and estimates of probe coverage across the genome indicate that the majority of all eutherian and avian/reptilian genes or regions of interest can be isolated using Uprobe. Future implementation of the universal probe design strategy will be used to create an expanded number of whole-genome probe sets that will encompass all vertebrate genomes.

[Supplemental material is available online at www.genome.org.]

The advent of large-scale DNA sequencing has led to the complete sequence of multiple vertebrates' genomes, including human (International Human Genome Sequencing Consortium 2001), mouse (Mouse Genome Sequencing Consortium 2002), rat (Rat Genome Project Sequencing Consortium 2004), and *Fugu* (Aparicio et al. 2002), as well as the ongoing and future efforts to sequence many more species (Couzin 2003). While complete genomic sequences represent the genetic information encoding a given species, a comprehensive understanding as to the function of all primary DNA sequence has yet to be achieved. One simple and powerful approach for detecting putative functional elements in vertebrate genomes is through interspecies sequence comparisons (Hardison 2000; Pennacchio and Rubin 2001). The power of interspecies sequence comparisons to detect putative functional elements is strongly correlated with the number of species and the divergence among the species included in the comparison (Boffelli et al. 2003; Margulies et al. 2003; Thomas et al. 2003). In addition, multiple species sequence comparisons promise to provide a means to correlate the modification of ancestral and/or the emergence of new functional elements associated with phenotypic innovations (Sidow 2002), as well a better understanding of genome evolution.

While whole-genome sequencing will yield a survey of many more vertebrates in the near future, targeted comparative mapping and sequencing is an efficient, rapid, and complementary approach for addressing directed biological questions in an even greater breadth of species (Gottgens et al. 2000; Chiu et al. 2002, 2004; Thomas et al. 2003). In addition, by providing high-quality sequence across specific regions of interest, targeted comparative mapping and sequencing can also supplement low coverage draft whole-genome sequence assemblies. Previously, as a means to facilitate targeted comparative physical mapping in placental mammals, we integrated the concept of "universal" sequence-tagged sites, which are PCR-based markers designed from sequences conserved between two or more species that can be used for mapping in multiple species (Venta et al. 1996; Lyons et al. 1997; Jiang et al. 1998), with overgo-hybridization probes, which are efficient at screening arrayed genomic libraries (Vollrath 1999). The resulting universal overgo-hybridization probes designed from sequences highly conserved in human and mouse were very efficient at isolating orthologous genomic segments from chimpanzee, baboon, cat, dog, cow, and pig bacterial artificial chromosome (BAC) libraries (Thomas et al. 2002). The use of these universal probes ultimately facilitated the direct comparison of megabases of orthologous sequence between these and other species (Thomas et al. 2003). Since the development and implementation of the universal probe design strategy, there has been significant expansion in the number of vertebrate BAC libraries. At the present time, at least 70 vertebrate genomic libraries are either available or under construction (see

**[3]Corresponding author.**
**E-mail jthomas@genetics.emory.edu; fax (404) 727-3949.**
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3066805. Article published online before print in December 2004.


15:166–173 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

http://bacpac.chori.org, http://www.benaroyaresearch.org/bri_investigators/amemiya/default.htm, http://www.genome.arizona.edu/, https://www.genome.clemson.edu/orders/, http:evogen.jgi.doe.gov/top_level/BAC.html, and http://hcgs.unh.edu) and are primarily accessible to clone isolation by hybridization-based screening. However, complementary genome-wide hybridization probe resources are not available for efficient screening of these libraries.

Here we report the implementation of the universal probe design concept across entire genomes for the creation of whole-genome universal probe sets that can be used to isolate orthologous chromosomal segments in specific evolutionary clusters of species. Based on a detailed retrospective analysis of prior universal probe hybridization results (Thomas et al. 2002) and new probe design algorithms, we have generated and experimentally validated hybridization-based whole-genome probe sets that can be used to effectively identify the vast majority of all eutherian (placental mammals) or avian/reptilian genes or regions of interest. To provide public access to this comparative mapping resource, we have created a Web site, Uprobe (http://uprobe.genetics.emory.edu), where simple queries can be used to identify, view, and download universal probe sequences from genes or regions of interest. Additional information is also accessible through the Uprobe Web site, including experimental protocols for using the universal probes, as well as all computational resources related to the creation of the whole-genome universal probe sets. Therefore, Uprobe provides a critical resource necessary for the efficient and widespread use of vertebrate BAC libraries for targeted comparative mapping and sequencing.

## Results

### Retrospective characterization of probe–target sequence hybridization

Previously, we reported the design and testing of 36-bp universal overgo-hybridization probes from highly conserved sequences between human and mouse, and their utility for screening chimpanzee, baboon, cat, dog, cow, and pig BAC libraries (Thomas et al. 2002). In that study, the success rate of the universal probes, defined as the fraction of probes that identified at least one BAC clone, was calculated for each species. However, the relationship between the hybridization outcome of each probe and the number of mismatches between the probe sequence and the sequence of the individual species (i.e., target sequence), was not determined. In order to better define the optimal criteria for the design of a whole-genome set of universal probes, a comprehensive retrospective analysis was performed to establish the number of mismatches tolerated between the probe and target sequence using newly available genomic sequence from the six species listed above. The results of this analysis are illustrated in Figure 1.

As expected, there was a clear correlation between the number of probe–target sequence mismatches and the presence of a positive hybridization signal. Probes with zero, three, or more than four mismatches with the target sequence were associated with positive hybridization signals 96%, 61%, and 26% of the time, respectively. We therefore conclude that the majority of individual 36-bp probes with three or fewer mismatches to a target sequence will result in a hybridization signal that is readily detectable using our BAC library hybridization protocol. Previously, orthologous BAC clones spanning regions of interest were isolated successfully using pools of probes spaced at ~30-kb intervals with an average success rate of >50% (Thomas et al. 2002). Thus, a whole-genome universal probe set comprised of probes expected to have three or fewer sequence mismatches with a specified target set of species would be an effective resource for isolating a gene or region of interest from one or a collection of BAC libraries.

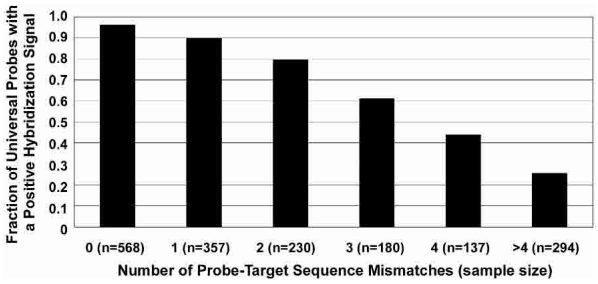



**Figure 1.** Retrospective analysis of universal probe–target sequence mismatches and hybridization signal. Overgo-hybridization probes ($n$ = 341) designed from human sequence and used to screen chimpanzee, baboon, cat, dog, cow, and pig BAC libraries (Thomas et al. 2002) were compared to newly available genomic sequence from each of these six target species. Local probe–target species sequence alignments were extracted from long-range human–target species genomic sequence alignments, and the number of mismatches recorded for each probe–target sequence comparison. The probe–target species sequence mismatch information was then combined with the probe hybridization results. A graph of the fraction of probe–target pairs that yielded a positive hybridization signal for a given number of probe–target sequence mismatches is indicated.

### Design and properties of a mammalian whole-genome universal probe set

In order to create a whole-genome probe set of universal probes for screening mammalian BAC libraries, we used available whole-genome alignments between human and mouse or human–mouse–rat to identify candidate probe sequences (human) likely to have three or fewer mismatches with the orthologous sequence in all other mammals. In order to accomplish this, modified versions of the overgo hybridization-probe selection program SOOP (Thomas et al. 2002) were created to design universal probes from either whole-genome pairwise or three-way alignments (see Methods). Since the number of substitutions per site between human and rodents is greater than that between human and most other placental mammals (Eizirik et al. 2001), we predicted that probes from human sequence with four or fewer mismatches with mouse would in most cases have three or fewer mismatches with the orthologous sequence in other placental mammals. In the case of dog, cat, cow, and pig, we were able to confirm that this prediction was in fact valid >75% of the time (see Methods). For the human–mouse–rat alignments, a more empirical scoring matrix and cutoff threshold were implemented (for details, see Methods) to design a similar set of universal probes. In both probe design methods, a score indicating the degree of conservation was assigned to each probe. Once probes were identified based on the above criteria, the sequence for each probe was classified as either unique or nonunique in the human genome as a means to predict whether or not a probe would likely identify a single locus in other mammals. While typically problematic for screening genomic libraries, nonunique probes were separated from unique probes and retained to facilitate the isolation of duplicated loci, such as gene families or segmental duplications. Comparison of the probe scores and sequence com-

position of the unique and nonunique probes did not reveal any major differences between these two probe types (data not shown). Finally, the universal probes designed by the two methods described above were merged to create a single mammalian universal probe set (FEB_2004_mammals_1) consisting of $n = 361{,}986$ unique probes and $n = 319{,}798$ nonunique probes.

We next sought to estimate the effective genome coverage of this probe set, i.e., the ability to isolate the orthologous genomic segment from another mammal to any given region of the human genome. To do so, we determined the number of unique probes in 200-kb intervals (50-kb slide) across the genome, assuming an optimal spacing of one probe every 30 kb (Fig. 2A). Assuming a probe success rate of at least 50%, the number of optimally spaced unique probes (as opposed to all unique probes) within the interval should therefore provide a conservative estimate as to the ability to isolate the orthologous region from a given BAC library. Using these criteria, we found that 95%, 86%, and 50% of all 200-kb intervals had at least one, three, or six optimally spaced unique probes, respectively. Overall, the average number of optimally spaced unique probes in each 200-kb interval was estimated to be 5.07 ± 2.12. With the exception of human chromosomes 19, X, and Y, 84% to 94% of all 200-kb intervals on a given chromosome had at least three optimally spaced unique probes (Supplemental Table 1). Human chromosomes 19, X, and Y had the lowest estimated probe coverage with just 67%, 69%, and 2% of all 200-kb intervals containing at least three optimally spaced unique probes, respectively. The low unique probe coverage observed on these chromosomes is likely due to a combination of factors, including overrepresentation of segmental duplications, gene families, and repetitive elements (Skaletsky et al. 2003; Grimwood et al. 2004) and lack of a rodent Y chromosome sequence (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Consortium 2004). Thus, this whole-genome universal probe set provides sufficient coverage for the isolation of mammalian BAC clones orthologous to the vast majority of the human genome.

## Experimental validation of the whole-genome mammalian universal probe set

To validate the probe design methods and estimate a probe success rate in different mammalian lineages, a sample of representative unique mammalian universal probes ($n = 95$) and one nonunique probe were selected for experimental validation. (A list and detailed description of the experimentally tested probes can be found in Supplemental Table 2.) Specifically, the probes selected for experimental validation had an average probe score (32.47 ± 0.42) between that of all unique probes (32.38 ± 0.43), and an optimal subset of probes spaced at 30-kb intervals (32.56 ± 0.44). The composition of the test set of probes based on their classification as either protein coding, UTR, intragenic, or intergenic was also similar to the entire set of unique probes (36.57%, 7.06%, 31.37%, and 25.00% versus 32.06%, 3.10%, 24.16%, and 40.68%). In addition, the selected universal probes were derived from 11 distinct chromosomal locations within the human genome that included a range of gene, GC, and repetitive element content (Supplemental Table 3).

The test set of universal probes was hybridized to a panel of nine BAC libraries (marmoset, galago, rabbit, bat, shrew, armadillo, elephant, wallaby, and platypus) selected to represent the major mammalian lineages (Novacek 1992; Nowak 1999; Murphy et al. 2001). The success rates, defined as the fraction of probes that detected at least one BAC clone, of the universal probes in the nine species are illustrated in Figure 2B. Among the placental mammals, the universal probes were associated with success rates between 58% (armadillo) and 95% (marmoset), suggesting that the probe design criteria were appropriate for effec-

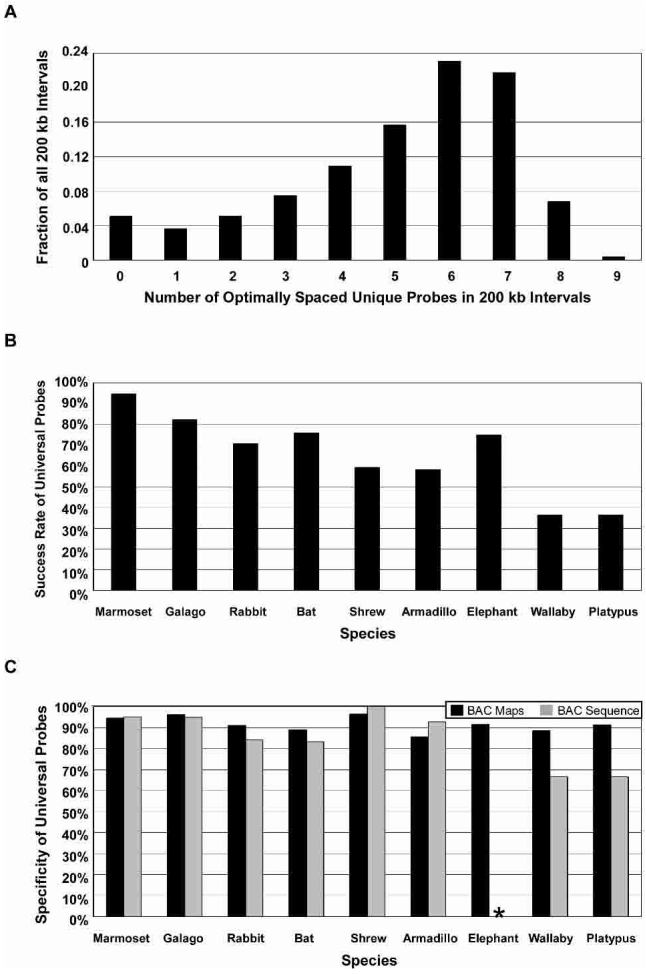



**Figure 2.** Characterization of the mammalian whole-genome universal hybridization probe set. (*A*) To estimate the fraction of the human genome associated with one or more mammalian universal probes, the number of unique probes selected using an optimal spacing parameter of 30 kb in all 200-kb windows (50-kb slide) across the genome, excluding intervals that contained a sequence gap in the human assembly, was determined. The result of that analysis was plotted as the fraction of all 200-kb intervals that included *N* number of optimally spaced unique probes. (*B*) To estimate the ability of universal probes to identify a BAC clone from mammalian libraries, a sample of $n = 96$ universal probes were hybridized to a panel of nine representative mammalian BAC libraries. The success rate, as defined by the percentage of universal probes that detected at least one positive BAC clone, for each of the nine species is indicated. (*C*) Specificity of the universal probes, as measured by physical mapping, and sequence analysis of individual isolated BAC clones is indicated. The number of BAC sequences included in the analysis for each species is as follows: marmoset, $n = 21$; galago, $n = 20$; rabbit, $n = 19$; bat $n = 12$; shrew, $n = 16$; armadillo $n = 14$; wallaby, $n = 12$; and platypus, $n = 9$. In the case of marmoset, shrew, and armadillo, probe specificity as measured by analysis of sequenced BAC clones was slightly higher than that of the probes themselves. Since multiple linked probes were included in the test set of probes, in these species many clones were positive for more than one probe and therefore facilitated a more accurate selection of orthologous versus nonorthologous clones for sequencing. *No BAC sequence is available for elephant.

tively screening BAC libraries from placental mammals. In contrast, a statistically significant lower success rate of 36% was observed for wallaby (marsupial) and platypus (monotreme) compared with the minimum success rate observed in placental mammals ($\chi^2$ test, $P < 0.05$). Thus, alignments that include additional species will likely be necessary to realize a similar probe success rate in marsupials and monotremes. It was also noted that, as expected, the probe score was correlated with the success or failure of the universal probes (Pearson correlation coefficient = 0.331) (Supplemental Fig. 1A).

An additional important property of the universal probes, beyond their ability to detect BAC clones, is specificity. We measured the specificity of the experimentally tested probes by two means. First, we used probe-content and restriction enzyme fingerprinting of BAC contigs to determine the percentage of probes that hybridized to a single location in a given species' genome (Fig. 2C). By this measure, at least 86% of the unique mammalian universal probes were single copy, and none showed a hybridization result expected for a common species-specific repetitive element. Second, sequence analysis was used to directly determine the percentage of orthologous BAC clones (Fig. 2C). By this measure, the mammalian universal probes were 83%–100% specific in placental mammals while 66% specific in both wallaby and platypus. Thus, the unique mammalian universal probes are highly specific and effective for isolating orthologous genomic regions from the BAC libraries of placental mammals.

### Design, characterization, and experimental validation of a bird/reptile whole-genome universal probe set

We next sought to implement the whole-genome probe design process toward the generation of an analogous universal probe set for screening bird and reptile BAC libraries. To do so, chicken–human whole genome alignments were used to identify candidate probe sequences (chicken) likely to have three or fewer mismatches with the orthologous sequence in other birds and reptiles. Since the divergence between chicken and human is significantly greater than that between chicken and other birds and reptiles (Kumar and Hedges 1998), a cutoff of four or fewer mismatches between chicken and human was selected. Using this criteria and subsequent classification of the probes as unique or nonunique in the chicken genome, a bird/reptile whole-genome set of universal probes (MAR_2004_birds/reptiles_1) consisting of $n = 73{,}720$ unique probes and $n = 36{,}197$ nonunique probes, was created. As with the mammalian whole-genome probe set, no other major differences were observed between the unique and nonunique probes (data not shown).

To characterize the bird/reptile whole-genome universal probe set, we estimated probe coverage across the chicken genome using the same methods described above for the mammalian whole-genome probe set. In this case, 91% of the 200-kb intervals had at least one unique probe, 64% had three or more optimally spaced unique probes, and 50% of the 200-kb intervals had at least four optimally spaced unique probes (Fig. 3A). The average number of optimally spaced unique probes per 200-kb interval was 3.52 ± 2.20. At the chromosome level, the fraction of 200-kb windows with three or more optimally spaced unique probes on the nonrandom chicken chromosome assemblies ranged from 56% (chromosome 22) to 94% (chromosomes 19 and 21). In contrast, only 27% of the intervals on the sequence assembly designated chicken chromosome "Unknown" had

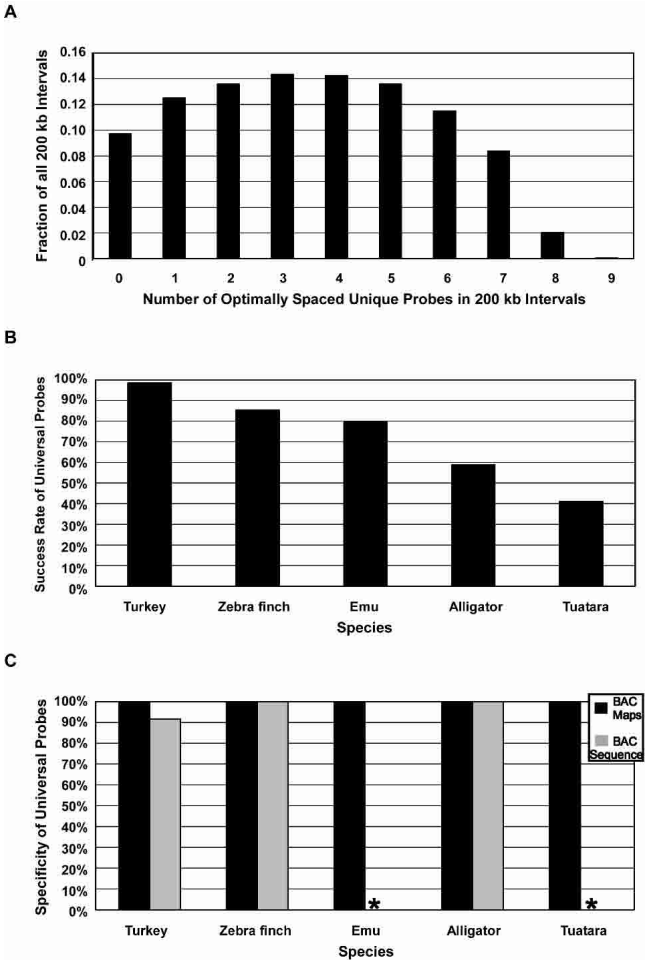



**Figure 3.** Characterization of the bird/reptile whole-genome universal hybridization probe set. (*A*) To estimate the fraction of the chicken genome associated with one or more bird/reptile universal probes, the number of unique probes selected using an optimal spacing parameter of 30 kb in all 200-kb windows (50-kb slide) across the genome, excluding intervals that contained a clone or centromere gap, was determined. The result of that analysis was plotted as the fraction of all 200-kb intervals that included *N* number of optimally spaced unique probes. (*B*) To estimate the ability of the universal probes to identify a BAC clone from avian/reptilian libraries, a sample of $n = 68$ universal probes was hybridized to a panel of five bird and reptile BAC libraries. The success rate, as defined by the percentage of universal probes that detected at least one positive BAC clone, for each of the five species is indicated. (*C*) Specificity of the universal probes as measured by physical mapping, and sequence analysis of individual isolated BAC clones is indicated. The number of BAC sequences included in the analysis for each species is as follows: turkey, $n = 12$; zebra finch, $n = 15$; and alligator, $n = 6$. *No BAC sequence is available for emu and tuatara.

three or more optimally spaced unique probes (Supplemental Table 4). Thus, the estimated coverage was lower than that for the mammalian universal probe set; however, it was anticipated that the evolutionary distance over which these probes would be effective would be significantly greater.

To test this hypothesis and experimentally validate the bird/reptile whole-genome universal probe set, we selected a set of representative unique probes for hybridization on a panel of bird and reptile BAC libraries. (A list and detailed description of the experimentally tested probes can be found in Supplemental Table 5.) The $n = 68$ selected probes had an average probe score

(93.96 ± 3.75) intermediate to that of all unique probes (92.79 ± 3.64) or an optimal subset of unique probes spaced every ~30 kb (94.17 ± 3.86). In addition, the selected probes were similar in base composition to that of all unique probes (75.82% putative exonic, 16.63% intragenic, and 7.56% intergenic for the test set versus 69.73%, 10.77%, and 19.50% for all unique probes). Finally, the selected probes were from eight distinct segments of the chicken genome that are at least partially orthologous to the regions chosen for mammalian universal probe testing (Supplemental Table 6).

The test set of universal bird/reptile probes were hybridized to a panel of BAC libraries from birds and reptiles (turkey, zebra finch, emu, alligator, and tuatara) that last shared a common ancestor with chicken between ~30 to 225 million years ago (Hedges and Poling 1999; van Tuinen and Hedges 2001; Dimcheff et al. 2002). The probe success rates varied from 98% in turkey to 41% in tuatara (Fig. 3B) and were correlated with estimated evolutionary distance from chicken and probe score (Pearson correlation coefficient = 0.243) (Supplemental Fig. 1B). Specificity of the bird/reptile probes was found to be at least 90% for all species based either on the probe-content and fingerprinted BAC contigs, or BAC sequences (Fig. 3C). Thus, the bird/reptile whole-genome universal probe set is effective for specifically isolating orthologous regions from BAC libraries of species that diverged as much as ~225 million years ago.

### Uprobe: Public access to whole-genome universal overgo probe sets

In order to provide public access to the universal whole-genome probe sets, a Web site, Uprobe (http://uprobe.genetics.emory.edu), was created. The primary purpose of the Uprobe Web site is to allow users the ability to identify universal probes for screening a given library, or collection of libraries, for a specific gene or region of interest. In order to accomplish this, we created a query page to a local database that includes both the universal probes, and the annotation and other positional information for the "reference" genome for each whole-genome probe set, i.e., human for mammals and chicken for birds/reptiles, from the UCSC Genome Browser (http://genome.ucsc.edu/; Karolchik et al. 2003). This allows the user to define a number of parameters, including chromosome location, gene name, GenBank accession number, or keyword to find universal probes of interest. In addition, the users can specify unique or nonunique probes, an optimal distance between probes, as well as a minimum probe score. As a result, users can customize the physical spacing between probes as well as the desired level of sequence divergence within the probe sequence to match their specific needs. For example, when the expected probe success rate is just ~50%, it has been our experience here and elsewhere (Thomas et al. 2002) that the hybridization of pools of at least five or more universal probes ideally spaced at ≤30-kb intervals is the most effective approach for the isolation of single genes or larger regions of interest. Thus, to maximize the utility of the universal probes, we strongly encourage users to employ similar set of criteria when using the Uprobe resource. Sequences for both primers necessary to generate the hybridization probe, and the full probe sequences can then be simply downloaded by the user, along with accompanying positional information. To verify the location of the universal probes and orient the user as to the position of the probes relative to one another as well as to their gene or region of interest, the location of selected probes can also be viewed on the UCSC Genome Browser via a link from Uprobe.

Along with the query interface, the Uprobe Web site presents the universal probe concept, how the probes are designed, statistical information on each whole-genome universal probe set, detailed protocols for using the universal probes, the complete set of tables included in Uprobe database, and the computer programs that created the universal probe sets. The Uprobe Web site therefore offers a comprehensive resource for universal overgo-hybridization probes that can be used in a simple and systematic manner to screen eutherian and avian/reptilian genomic BAC libraries.

## Discussion

Comparative sequence analysis in vertebrates is becoming a powerful and common method for inferring function from primary genomic sequence (Mouse Genome Sequencing Consortium 2002; Rat Genome Project Sequencing Consortium 2004). Comparative physical mapping and sequencing of targeted intervals can provide extensive information on putative functional elements and the molecular evolution of genes or regions of interest (Gottgens et al. 2000; Loots et al. 2000; Chiu et al. 2002, 2004; Thomas et al. 2003). Physical mapping also yields DNA clones that can be used in a variety of experiments, such as the creation of transgenics (Antoch et al. 1997; Barton et al. 2001), or cytogenetic mapping (Kirsch et al. 2000; Eder et al. 2003). Historically, targeted comparative physical mapping has been limited by the availability of genomic libraries. Programs aimed at increasing the number of genomic BAC libraries from vertebrates have successfully addressed this limitation (http://www.genome.gov/10001844, and http://www.nsf.gov/bio/pubs/awards/bachome.htm), such that now more than seventy BAC libraries are available or in progress. Thus, the practical limitation for using this extensive BAC library resource is the ability to efficiently identify genomic clones containing genes or regions of interest.

Overgo-hybridization probes provide a simple, scalable, and specific means for BAC library screening (The International Human Genome Mapping Consortium 2001; Gregory et al. 2002). By using available whole-genome sequences and alignments, we have created whole-genome sets of universal overgo-hybridization probes for screening BAC libraries from clusters of related species. In particular, we have established a whole-genome probe set for screening BAC libraries from placental mammals. Using both computational and direct experimental methods, we estimate that this mammalian probe set can be used to efficiently isolate genomic clones orthologous to the vast majority of the human genome from any of the current ($n$ = ~40), or future eutherian genomic libraries. An analogous set of hybridization probes was created and experimentally validated for screening bird and reptile genomic libraries. While the estimated genome coverage of the bird/reptile probe set was lower than that of the corresponding mammalian probe resource, the bird/reptile probes were able to effectively screen libraries from species that diverged ~225 million years ago (Hedges and Poling 1999), versus the ~108 million years (Murphy et al. 2001) observed for the mammalian probes. Therefore, both whole-genome probe sets provide a general resource for selecting pools of physically linked probes, that in aggregate have been shown to be a valuable

tool for isolating orthologous genes or regions of interest from multiple species in parallel (Thomas et al. 2002).

As the number of vertebrate genomes that are sequenced increases, one might expect that the use of BAC libraries would diminish. However, we believe that as more genomes are sequenced, there will be a concomitant increase in the utility of targeted comparative mapping and sequencing of genes of regions of interest in clusters of closely related species. Specifically, targeted comparative mapping and sequencing could be used as a means to correlate observed phenotypic variation with sequence variation between species, with access to large-insert genomic clones providing a critical experimental resource needed to test inferred genotype-phenotype correlations. Thus, we plan to implement and improve the strategies outlined here with future whole-genome sequences and alignments in an effort to expand and enhance the described universal probe resource to include all clusters of vertebrates for which BAC libraries are available.

## Methods

### Design and selection of universal hybridization probes

Two approaches were taken for designing whole-genome probe sets. In the first, modified versions of the probe selection program SOOP (Thomas et al. 2002), sooper.xml and sooper_xml_v2, were used to select 36-bp probes using human–mouse and chicken–human pairwise alignments. A percent sequence identity of >88% (equivalent to four or fewer mismatches over the length of the probe) was used as the cutoff value for probe selection. In the second approach, human–mouse–rat alignments were the basis of probe selection using a newly developed program, multi_soop. To optimize the selection of the best probes and fully utilize the information that a third species added to the alignments, a new probe scoring scheme was implemented based on an empirically derived scoring matrix specific for human–mouse–rat alignments. Determination of the matrix values was done as follows. A set of n = 2863 36-bp sequences with fewer than seven mismatches between human and mouse and for which corresponding orthologous sequences were available in rat, cow, dog, cat, and pig was identified. Each pattern ($n$ = 5) of matches and mismatches at a single nucleotide position between human–mouse–rat was assigned a value based on the frequency at which the human nucleotide in each of the five patterns was identical to the orthologous cow, dog, cat, and pig nucleotide. A probe "score" was then calculated by summing the matrix values derived from the human–mouse–rat alignment for each position in the probe. A multi_soop score predicted to include >93% of all probes with zero, one, two, or three mismatches between human and mouse and exclude 60% of all probes with four or more mismatches was set as the cutoff value for probe selection. More details on the scoring matrix are available at http://uprobe.genetics.emory.edu/direction.html.

From the complete set of 36-bp sequences with fewer than seven mismatches between human and mouse and for which corresponding orthologous sequences were available in rat, cow, dog, cat, and pig, the $n$ = 1739 that had four or fewer human–mouse mismatches were extracted and used to calculate the fraction of dog, cat, cow, and pig orthologous sequences that also had four or fewer mismatches.

All selected candidate probes were compared back to their genome of origin, either human (human–rodent alignments) or chicken (chicken–human alignments), with MEGABLAST (megablast -t 16 -N 2 -W 11 -e 0.6 -F F -D 3) (Zhang et al. 2000) to determine if it was unique or nonunique. Probes were classified as unique if they had a single identical match to the genome of origin, no other matches with a bit score >40, and fewer than five matches with a bit score >36. Probes were classified as nonunique if they had, in addition to the expected identical match to the genome of origin, at least one other match with a bit score >40, or five or more matches with bit scores >36.

The computer programs developed and used in the probe design process can be downloaded from http://uprobe.genetics.emory.edu/scripts.php.

### Whole-genome alignments and annotation

Whole-genome alignments used for universal probe design were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/; Karolchik et al. 2003) and included: the April 2003 human genome assembly (UCSC version hg15) and the February 2003 mouse genome assembly (UCSC version mm3; axtTight) alignments (Schwartz et al. 2003); the human (hg15)-mouse (mm3)–rat (January 2003, UCSC version rn2) MULTIZ alignments (Blanchette et al. 2004); and the chicken (February 2003, UCSC galGal2)–human (July 2003, UCSC version hg16) alignments. Probes were classified based on the sequence of origin, i.e., coding, intragenic, etc, using whole-genome human (hg15) and chicken (galGal2) annotation imported from the UCSC Genome Browser (http://genome.ucsc.edu/; Karolchik et al. 2003).

### Experimental validation of universal probes

Experimental validation of the whole-genome universal probe sets was done by hybridizing samples of unique probes on a panel of BAC libraries following the methods described in Thomas et al. (2002). (A single nonunique probe originating from a physically linked duplicated locus, laboratory name 33e9, was also tested [Supplemental Table 2].) Briefly, sets of $n$ = 48 universal overgo-hybridization probes (comprising two complementary 22-mers with an 8-bp overlap and radio-labeled by a primer-extension reaction with Klenow in the presence of [$\alpha^{32}$P]dATP and [$\alpha^{32}$P]dCTP) (Vollrath 1999) were hybridized to multiple libraries in parallel with a single set of hybridization and washing conditions. Specifically, hybridization overnight at 58°C in Church buffer, followed by a 15-min wash at 58°C in 2× SSC, 0.1% SDS, and then washes for 30 min each at 58°C in 1.5× SSC 0.1% SDS and 1.0× SSC, 0.1% SDS. Positive BAC clones were scored using Comboscreen (Jamison et al. 2000) and single colonies rearrayed for generation of secondary filters using a BioGrid (BioRobotics), as well as for restriction-enzyme fingerprinting (Marra et al. 1997). Restriction-enzyme fingerprints for each BAC clone were imported into IMAGE (www.sanger.ac.uk/Software/Image/) and contigs assembled with FPC (Soderlund et al. 2000). Probe-content information for the individual BAC clones was generated by hybridization of the universal probes to secondary filters and then viewed and edited with SEGMAP (Green and Green 1991). A complete set of detailed protocols for use of the universal probes can be obtained from http://uprobe.genetics.emory.edu/process.php.

### Quantification of universal probe specificity

The combination of BAC clone probe-content maps and restriction-enzyme fingerprint contigs was used to assess whether or not each probe identified a single locus within a given species (i.e., single copy or non-single copy). Probes were designated single copy if the number of clones positive for a given probe was less than 3.5× the expected depth of library coverage, and if greater than two-thirds of all clones positive for a given probe were within a single restriction-enzyme fingerprint contig (ex-

cluding singletons). Probes that did not meet these criteria were designated non-single copy. In addition, the probe-content and restriction-enzyme fingerprint contigs were used for the selection of a set of representative and minimally overlapping BAC clones for sequencing.

BAC clones were full-shotgun sequenced at the National Institutes of Health (NIH) Intramural Sequencing Center and submitted to GenBank immediately after assembly. The sequence from each clone was then evaluated to determine whether or not it was orthologous to the targeted interval (Thomas et al. 2002). Specifically, BAC clones were considered orthologous in cases in which the observed and expected probe content and sequence overlaps with neighboring clones were consistent, and comparison of the assembled BAC clone sequence to the reference human or chicken genomic sequence revealed a pattern of alignments in the noncoding and coding segments of one or more genes consistent with the degree to which the species being compared have diverged. BAC clones classified as nonorthologous typically did not meet any of these criteria. In particular, BAC clones were designated nonorthologous in cases in which comparison to the reference human or chicken genomic sequence revealed either no alignments, a few random alignments, or alignments restricted to the coding segments of a single gene.

### BAC libraries

Arrayed BAC library filter sets and individual BAC clones were obtained from: BACPAC Resources Center (http://bacpac.chori.org/) (galago [*Otolemur garnetti*, CHORI-256], marmoset [*Callithrix jacchus*, CHORI-259], turkey [*Meleagris gallopavo*, CHORI-260], armadillo [*Dasypus novemcinctus*, VMRC-5], bat [*Rhinolophus ferrumequinum*, VMRC-7], elephant [*Loxodonta africana*, VMRC-15], and rabbit [*Oryctolagus cuniculus*, LBNL-1]); Benaroya Research Institute at Virginia Mason (http://www.benaroyaresearch.org/bri_investigators/amemiya/default.htm) (alligator [*Alligator Mississippiensis*, VMRC-8], tuatara [*Sphenodon punctatus*, VMRC-12], and emu [*Dromaius novaehollandiae*, VMRC-16]); Arizona Genomics Institute (http://www.genome.arizona.edu/) (wallaby [*Macropus eugeni*, ME_Kba] and zebra finch [*Taeniopygia guttata*, TG_Ba]); and Clemson University Genomics Institute (https://www.genome.clemson.edu/) (platypus [*Ornithorhynchus anatinus*, OA_Bb] and shrew [*Sorex araneus*, SA_Ba]). Note, the elephant, armadillo, and bat libraries were constructed at the Benaroya Research Institute at Virginia Mason and the rabbit library at the Lawrence Berkeley National Laboratory.

## Acknowledgments

We acknowledge the contributions of the members of the NISC Comparative Sequencing Program, especially E.D. Green, R.W. Blakesley, A.C. Young, B. Maskeri, and J.C. McDowell. We also thank P. DeJong, C. Amemiya, R. Wing, J. Tompkins, J.F. Cheng, and S. Edwards for construction and access to the BAC libraries used in this report; C. Tsui and S. Pittard for computational support; and R. Grier for technical support. This work was supported by the NIH BAC Resource Network grant MH068185.

## References

Antoch, M.P., Song, E.J., Chang, A.M., Vitaterna, M.H., Zhao, Y., Wilsbacher, L.D., Sangoram, A.M., King, D.P., Pinto, L.H., and Takahashi, J.S. 1997. Functional identification of the mouse circadian Clock gene by transgenic BAC rescue. *Cell* **89:** 655–667.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301–1310.

Barton, L.M., Gottgens, B., Gering, M., Gilbert, J.G., Grafham, D., Rogers, J., Bentley, D., Patient, R., and Green, A.R. 2001. Regulation of the stem cell leukemia (SCL) gene: A tale of two fishes. *Proc. Natl. Acad. Sci.* **98:** 6747–6752.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299:** 1391–1394.

Chiu, C.H., Amemiya, C., Dewar, K., Kim, C.B., Ruddle, F.H., and Wagner, G.P. 2002. Molecular evolution of the HoxA cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci.* **99:** 5492–5497.

Chiu, C.H., Dewar, K., Wagner, G.P., Takahashi, K., Ruddle, F., Ledje, C., Bartsch, P., Scemama, J.L., Stellwag, E., Fried, C., et al. 2004. Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res.* **14:** 11–17.

Couzin, J. 2003. Sequencers examine priorities. *Science* **301:** 1176–1177.

Dimcheff, D.E., Drovetski, S.V., and Mindell, D.P. 2002. Phylogeny of Tetraoninae and other galliform birds using mitochondrial 12S and ND2 genes. *Mol. Phylogenet. Evol.* **24:** 203–215.

Eder, V., Ventura, M., Ianigro, M., Teti, M., Rocchi, M., and Archidiacono, N. 2003. Chromosome 6 phylogeny in primates and centromere repositioning. *Mol. Biol. Evol.* **20:** 1506–1512.

Eizirik, E., Murphy, W.J., and O'Brien, S.J. 2001. Molecular dating and biogeography of the early placental mammal radiation. *J. Hered.* **92:** 212–219.

Gottgens, B., Barton, L.M., Gilbert, J.G.R., Bench, A.J., Sanchez, M.-J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M., et al. 2000. Analysis of vertebrate *SCL* loci identifies conserved enhancers. *Nat. Biotechnol.* **18:** 181–186.

Green, E.D. and Green, P. 1991. Sequence-tagged site (STS) content mapping of human chromosomes: Theoretical considerations and early experiences. *PCR Methods Appl.* **1:** 77–90.

Gregory, S.G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C.E., Evans, R.S., Burridge, P.W., Cox, T.V., Fox, C.A., et al. 2002. A physical map of the mouse genome. *Nature* **418:** 743–750.

Grimwood, J., Gordon, L.A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., et al. 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428:** 529–535.

Hardison, R.C. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16:** 369–372.

Hedges, S.B. and Poling, L.L. 1999. A molecular phylogeny of reptiles. *Science* **283:** 998–1001.

The International Human Genome Mapping Consortium. 2001. A physical map of the human genome. *Nature* **409:** 934–941.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Jamison, D.C., Thomas, J.W., and Green, E.D. 2000. ComboScreen facilitates the multiplex hybridization-based screening of high-density clone arrays. *Bioinformatics* **16:** 678–684.

Jiang, Z., Priat, C., and Galibert, F. 1998. Traced orthologous amplified sequence tags (TOASTs) and mammalian comparative maps. *Mamm. Genome* **9:** 577–587.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.

Kirsch, I.R., Green, E.D., Yonescu, R., Strausberg, R., Carter, N., Bentley, D., Leversha, M.A., Dunham, I., Braden, V.V., Hilgenfeld, E., et al. 2000. A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nat. Genet.* **24:** 339–340.

Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Lyons, L.A., Laughlin, T.F., Copeland, N.G., Jenkins, N.A., Womack, J.E., and O'Brien, S.J. 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nat. Genet.* **15:** 47–56.

Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7:** 1072–1084.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W., et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294:** 2348–2351.
Novacek, M.J. 1992. Mammalian phylogeny: Shaking the tree. *Nature* **356:** 121–125.
Nowak, R.M. 1999. *Walker's mammals of the world*. The Johns Hopkins University Press, Baltimore, MD.
Pennacchio, L.A. and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2:** 100–109.
Rat Genome Project Sequencing Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Sidow, A. 2002. Sequence first: Ask questions later. *Cell* **111:** 13–16.
Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423:** 825–837.
Soderlund, C., Humphray, S., Dunham, A., and French, L. 2000. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10:** 1772–1787.
Thomas, J.W., Prasad, A.B., Summers, T.J., Lee-Lin, S.Q., Maduro, V.V., Idol, J.R., Ryan, J.F., Thomas, P.J., McDowell, J.C., and Green, E.D. 2002. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* **12:** 1277–1285.
Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
van Tuinen, M. and Hedges, S.B. 2001. Calibration of avian molecular clocks. *Mol. Biol. Evol.* **18:** 206–213.
Venta, P.J., Brouillette, J.A., Yuzbasiyan-Gurkan, V., and Brewer, G.J. 1996. Gene-specific universal mammalian sequence-tagged sites: Application to the canine genome. *Biochem. Genet.* **34:** 321–341.
Vollrath, D. 1999. DNA markers for physical mapping. In *Genome analysis: A laboratory manual: Mapping genomes,* Vol. 4. (eds. B. Birren et al.), pp. 187–215. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214.

## Web site references

http://bacpac.chori.org; BACPAC Resources Center.
http://www.benaroyaresearch.org/bri_investigators/amemiya/default.htm; Amemiya Laboratory.
http://www.genome.arizona.edu/; Arizona Genomics Institute.
https://www.genome.clemson.edu/; Clemson Genomics Institute.
http://genome.ucsc.edu/; UCSC Genome Browser.
http://evogen.jgi.doe.gov/top_level/BAC.html; JGI Evolutionary Genomics.
http://hcgs.unh.edu; Hubbard Center for Genome Sciences.
http://uprobe.genetics.emory.edu, Uprobe.
http://www.genome.gov/10001844; NIH BAC Resource Network.
http://www.nsf.gov/bio/pubs/awards/bachome.htm; National Science Foundation–funded BAC libraries.

*Received July 27, 2004; accepted in revised form September 16, 2004.*