Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Genomics 87 (2006) 181 – 190
GENOMICS
www.elsevier.com/locate/ygeno

Method

Utilization of a zebra finch BAC library to determine the structure of an avian androgen receptor genomic region ☆

Meizhong Luo [a,b,*], Yeisoo Yu [a], HyeRan Kim [a], Dave Kudrna [a], Yuichiro Itoh [c], Robert J. Agate [c], Esther Melamed [c], José L. Goicoechea [a], Jayson Talag [a], Christopher Mueller [a], Wenming Wang [a,1], Jennifer Currie [a], Nicholas B. Sisneros [a], Rod A. Wing [a,*], Arthur P. Arnold [c,*]

[a] Department of Plant Sciences, Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA
[b] College of Life Sciences and Technology, Huazhong Agricultural University, Wuhan 430070, China
[c] Department of Physiological Science, University of California, Los Angeles, CA 90095, USA

Received 17 May 2005; accepted 7 September 2005
Available online 29 November 2005

Abstract

The zebra finch (Taeniopygia guttata) is an important model organism for studying behavior, neuroscience, avian biology, and evolution. To support the study of its genome, we constructed a BAC library (TG__Ba) using DNA from livers of females. The BAC library consists of 147,456 clones with 98% containing inserts of an average size of 134 kb and represents 15.5 haploid genome equivalents. By sequencing a whole BAC, a full-length androgen receptor open reading frame was identified, the first in an avian species. Comparison of BAC end sequences and the whole BAC sequence with the chicken genome draft sequence showed a high degree of conserved synteny between the zebra finch and the chicken genome.

Keywords: Androgen receptor; BAC library; Bird; Genomics; Zebra finch

The order Passeriformes in birds comprises approximately 5300 species, including approximately 4000 oscine songbirds. Songbirds have played a disproportionate role in the development of numerous subdisciplines of biology, probably because they are often conspicuous, diurnal species that can be observed readily in the field and are also tractable for laboratory studies. Therefore much is known about the behavior, breeding and population biology, ecology, and neurobiology of songbirds. The development of modern genetic resources for investigations of songbirds will greatly enhance genetic and molecular approaches to the study of songbirds in areas such as population genetics, ecology, and behavioral neurobiology [1].

Songbirds receive their group name from the complex, learned song which is part of male courtship and territorial displays. The species studied most in the laboratory is the zebra finch (Taeniopygia guttata, family Estrildidae), a native of Australia, which has become a popular cage bird because it breeds easily in captivity. Zebra finch males copy the song of their father during an early critical period of development. The discovery of a discrete brain circuit controlling song [2] made this system an important model for studying the neural basis of learning, sexual differentiation of the brain, adult neurogenesis, the effects of gonadal hormones on brain circuits, seasonal changes in the brain, auditory processing and sensory–motor integration, the evolution of brain circuits, and the neural basis of behavior [3].

The genomics approach has been shown to be powerful in uncovering genome organization, gene function, and evolution. Physical maps of many organisms such as human and chicken

☆ Sequence data from this article have been deposited with the GenBank Data Library under Accession Nos. CW991863 through CW991926 and CZ167565 to CZ167566 for the BAC end sequences, AC153487 for the BAC 319A15 sequence, AY847476 for the AR 5′RACE product, and BK005685 for our predicted zebra finch AR cDNA (pzfAR).

* Corresponding authors. M. Luo is to be contacted at Department of Plant Sciences, Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA. R.A. Wing, fax: +1 520 621 1257.
E-mail addresses: mluo@ag.arizona.edu (M. Luo), rwing@genome.arizona.edu (R.A. Wing), arnold@ucla.edu (A.P. Arnold).
[1] Current address: Biotechnology Institute, University of Maryland, College Park, MD 20742, USA.

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.09.005

[4,5] have been built. Genomes of many organisms including human and mouse [6,7] have been sequenced. Many important genes have been cloned through positional cloning strategies [8,9]. In the songbird Agelaius phoeniceus, cosmid sequencing was used to reveal the structure of Mhc Class II gene clusters [10].

However, although the zebra finch is important as a model system, the study of its genome lags far behind because of the lack of large-insert libraries of genomic DNA and the absence of a physical map of the genome. The recently released draft sequence of the chicken genome [11] serves as a reference sequence for birds. Because the chromosome structure of birds, including zebra finches, appears to have undergone relatively little rearrangement during avian evolution [12], it is likely that many of the syntenic relationships found in chickens will be conserved in zebra finches. Further study of zebra finch biology, however, requires much more genomic information than is currently available. A BAC library is the first important step toward further examination of the zebra finch genome. Moreover, the genomic information from the zebra finch will allow interesting comparisons to chicken and other species to accelerate understanding of the evolution of birds and other vertebrates.

High-quality large-insert genomic DNA libraries are critical tools for physical mapping, positional cloning, and genome sequencing. The BAC (bacterial artificial chromosome) cloning system is the predominant system for large insert genomic DNA cloning of all organisms [13–17]. To support the genomic study of the zebra finch, we constructed a BAC library using DNA from livers of females. By completely sequencing a BAC clone, we identified an open reading frame for the full-length androgen receptor, a gene that has resisted cloning in birds. Comparison of the BAC end sequences and the whole BAC sequence with the chicken genome draft sequence revealed a high degree of synteny between the zebra finch and the chicken genome.

Results

BAC library construction and insert size analysis

A BAC library was constructed for zebra finch using DNA from livers of females. Females were chosen because their genome contains both sex chromosomes (ZW) in contrast to the homogametic male genome (ZZ). The restriction enzyme HindIII was used for preparation of both vector and genomic DNA fragments. The library consists of 147,456 clones stored in 384 × 384-well microtiter plates. Of 348 clones randomly picked from the library, more than 98% contain inserts (average insert size 134 kb) and 79.3% contain inserts larger than 100 kb. According to the reported zebra finch haploid genome size of 1250 Mb, the same as that of chicken (http://www.genomesize.com), the library provides a coverage of 15.5 genome equivalents.

BAC library screening with gene-specific probes

To demonstrate the utility of the BAC library, we screened the library with 10 gene-specific probes. Putative positives of the first screening were verified by colony or Southern hybridization. The hybridization results are listed in Supplemental Table 1. The 10 probes each yielded 2 to 24 positive clones with an average of 14 per probe.

DNA analysis of the 15 putative positive androgen receptor (AR46) BAC clones of the first screening showed an average insert size of 144 kb (Fig. 1). From these 15 clones, 9 were confirmed by Southern hybridization with strong signals. The average insert size of these 9 clones is 150 kb and all have different NotI restriction patterns, indicative of high quality and representative cloning of this gene region. The remaining 6 showed only weak signals. These clones could be false positives from hybridization to related nuclear receptor family genes or positives containing only very short probe sequences at the ends of the inserts. These clones were not further analyzed in this study.

BAC end sequencing

To gain sequence information about the zebra finch genome, we sequenced both ends of the positive clones of the Hat3, BDNF, and AR46 probes (Supplemental Table 1). About 94% of BAC ends were sequenced successfully, with an average high-quality base pair number of 533 bp (GenBank CW991863 through CW991926 and CZ167565 to CZ167566). We performed BLAST searches of these BAC end sequences to the chicken genome and found that most matched to the chicken Hat3, BDNF, and AR orthologous regions, respectively (Fig. 2), although some of them also matched to other regions nonspecifically. The probe sequences mapped exclusively to their respective orthologous regions of the chicken genome. Fig. 2 does not show the positions corresponding to the probe sequences because these probe sequences are heterologous cDNA sequences to chicken and identification of the corresponding exons on the chicken genome draft sequence is not conclusive. Of 15 Hat3 positive clones, 6 clones matched to the chicken Hat3 orthologous region at both ends, 5 at one end, and 4 at neither end (match rate 56.7%). For 13 BDNF positive clones, 5 clones matched to the chicken BDNF orthologous region at both ends, 6 at one end, and 2 at neither end (match rate is 61.5%). For the 9 AR46 positive clones, 5 clones matched to the chicken AR orthologous region at both ends and 3 at one end. One clone, BAC 319A15 (Fig. 1), that had been selected for FISH mapping to metaphase chromosomes [12] and whole BAC sequencing (see below), did not match to this region at either end. The match rate is 72.2%. The average match rate at the three regions is 62.2%. This value may reflect some sequence

Table 1
CpG islands in the zebra finch BAC 319A15 sequence

CpG island	Start/end	Size (bp)	GC%
1	32,094/34,527	2434	74.28
2	48,349/48,436	88	69.32
3	117,672/117,848	177	64.41
4	123,416/124,766	1351	74.91
5	125,406/125,581	176	68.18

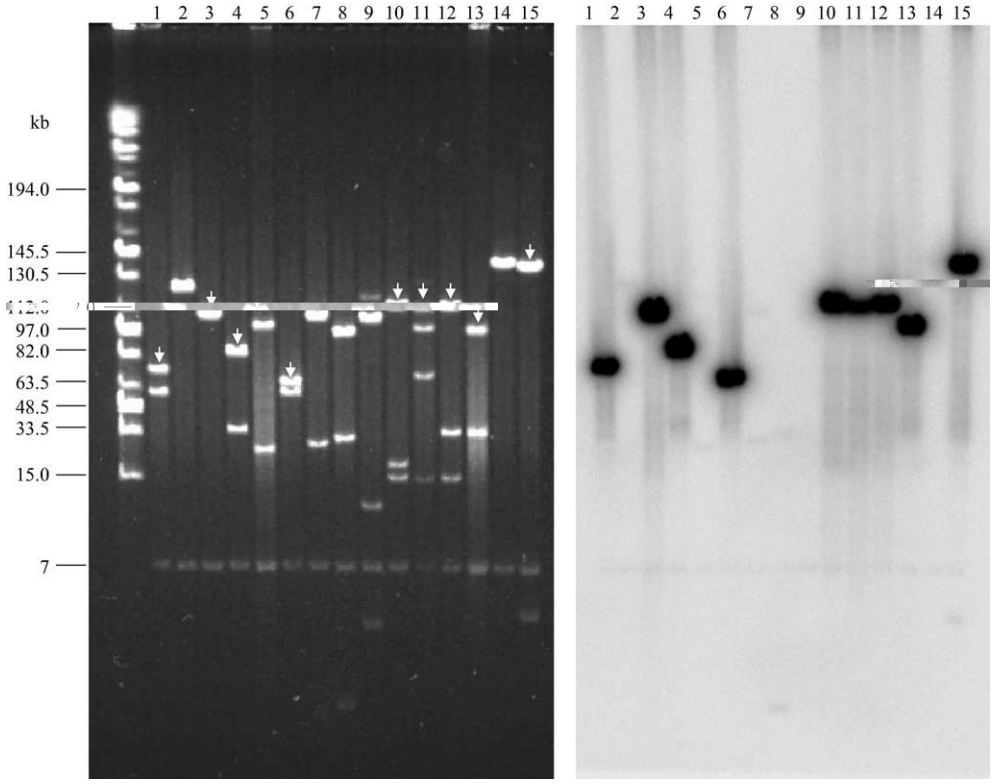

Fig. 1. DNA gel analysis and Southern hybridization of the zebra finch AR46 putative positive clones. Left: DNA of the zebra finch AR46 putative positive clones screened out from the BAC library was digested with NotI, separated on a 1% agarose CHEF gel at 5- to 15-s linear ramp time, 6 V/cm, 14°C in 0.5× TBE buffer for 16 h, and stained with ethidium bromide. The arrows indicate the bands recognized by the probe shown on the right. The common band is the vector. The marker used is PFG midrange I. Right: The same gel as on the left was blotted to a nylon membrane and hybridized with the AR46 probe. The numbers represent the addresses of the BAC clones analyzed: 1, 38L19; 2, 45B03; 3, 03O13; 4, 05P11; 5, 47M24; 6, 53O01; 7, 118P21; 8, 137C03; 9, 119C04; 10, 180L16; 11, 178I11; 12, 228D16; 13, 319A15; 14, 299M22; 15, 346M06.

divergence of the two genomes in noncoding regions. In general, however, these data indicate that the zebra finch genome is highly syntenic to the chicken genome and end sequencing of our BAC library can be used in making a zebra finch comparative genome map.

Sequencing of BAC 319A15 containing the androgen receptor gene

Because androgens play important roles in avian reproduction, there has been considerable interest in identifying the cDNA encoding the avian androgen receptor(s). To date, however, a full-length cDNA has not been described. Initial cloning isolated partial cDNAs for canary [18] (GenBank L25901) and zebra finch (AR46 [19]; 1394 bp; GenBank AF532914). Using 5′-RACE, we identified 447 bp additional sequence 5′ of AR46 (GenBank AY847476) and confirmed using RT-PCR that the two sequences are expressed as a single transcript. The 150 bp at the 5′ end of the RACE product contained 84.7% GC, and attempts to isolate more 5′ cDNA sequence via RACE failed. The middle of the 1802-bp contig formed by AR46 and the RACE product (called czfAR here) aligns well with the 3′ end of mammalian androgen receptors (e.g., GenBank mouse cDNA M37890) and with the predicted partial chicken AR (GenBank XM420163, 89% homology, czfAR bp 417–1430). However, the 5′ half of the mouse androgen receptor (~1500 bp) does not align by BLAST to any sequence in the chicken genome, suggesting that there is significant divergence of avian and mammalian AR cDNAs at the 5′ end. To obtain further information about the 5′ end of the zebra finch AR, we sequenced BAC 319A15, which was determined to contain at least part of the AR gene by probing Southern blots of BACs using both AR46 (Fig. 1) and a probe encoding the 5′ end of czfAR (not shown). BAC 319A15 was subsequently mapped by FISH to a zebra finch microchromosome homologous to the distal tip of chicken chromosome 4p [12].

The BAC 319A15 sequence (GenBank AC153487) contains 136,080 bp, which corresponds well to its size on the agarose gel (Fig. 1, left, lane 13). The sequence assembly was confirmed by comparing the in silico restriction patterns with the real restriction enzyme digestion patterns of BAC 319A15 for NotI (Fig. 1, left, lane 13) and HindIII, BamHI, XhoI, SwaI, SalI, and PvuI (data not shown). The average GC content of the BAC sequence is 50% and the GC content in 400-bp windows varies from 30 to 82% (Fig. 3A). Five CpG islands with minimum length >50 bp and GC content >64% were found (Fig. 3B). CpG islands were usually found at the 5′ ends of genes [20]. The positions, sizes, and GC content of these five CpG islands are listed in Table 1.

We searched the BAC sequence for genes with both GENSCAN (http://genes.mit.edu/GENSCAN.html) and FGE-

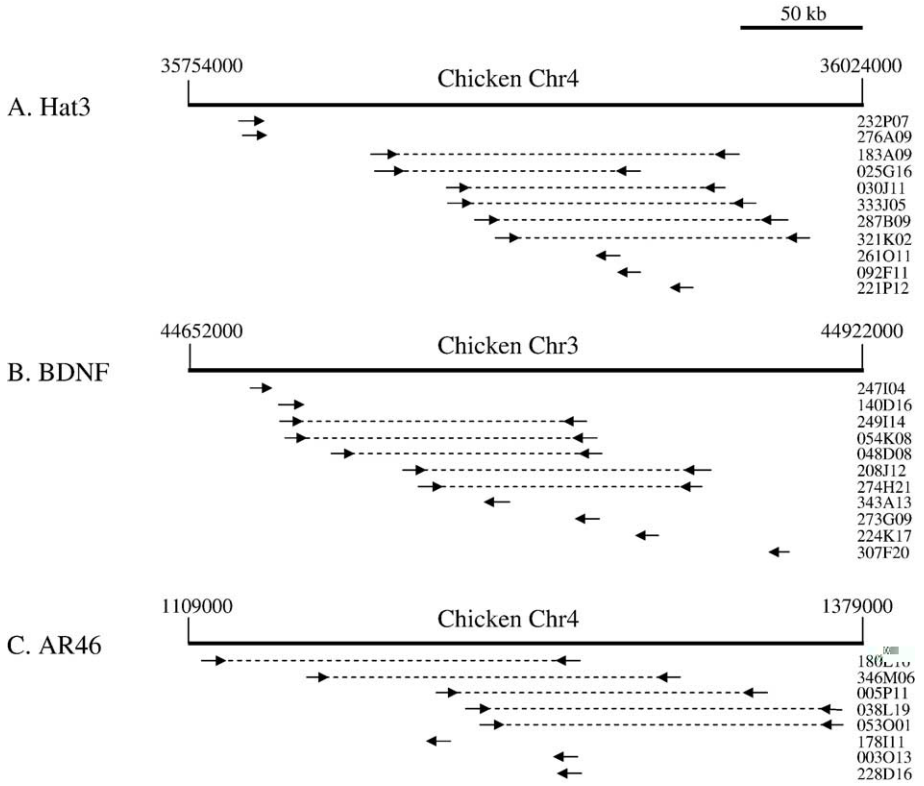

Fig. 2. Mapping of the end sequences of the positive BAC clones of the probes (A) Hat3, (B) BDNF, and (C) AR46 to the corresponding regions of the chicken genome draft sequence. The solid lines indicate the corresponding regions of the chicken genome draft sequence. The chromosome and genomic (nucleotide) locations are labeled above and at the ends, respectively. The arrows represent the locations and directions of the BAC end sequences. The broken lines are used to link the two end sequences of the same BAC clones. The clone addresses (names) are listed at the right side of the end sequences.

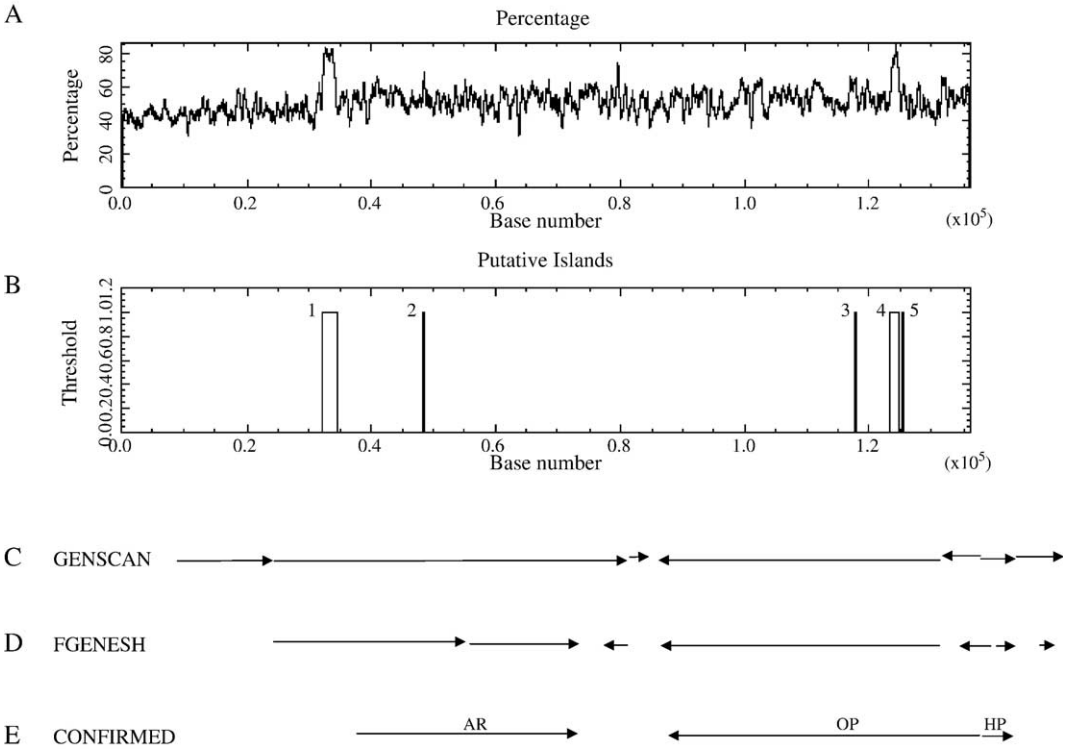

Fig. 3. Features of the zebra finch BAC 319A15 sequence. (A) GC content in 50-bp length and 400-bp windows. (B) CpG islands. Both A and B were determined with the online program www.ebi.ac.uk/emboss/cpgplot. (C) Genes predicted with GENSCAN (http://genes.mit.edu/GENSCAN.html). (D) Genes predicted with FGENESH (http://www.softberry.com/berry.phtml). (E) Genes confirmed by cDNA matches. AR, androgen receptor; OP, human oligophrenin 1 homolog; HP, human hypothetical protein MGC21416 homolog.

Table 2
Structural organization of the zebra finch BAC 319A15 sequence

Class	GENSCAN			FGENESH			Confirmed		
	Number	Total length (bp)	% of length	Number	Total length (bp)	% of length	Number	Total length (bp)	% of length
Genes	7	116,814	85.84	7	86,928	63.88	3	82,445	60.59
Exons	52	8422	6.19	45	7419	5.45	34	5341	3.92
Introns	45	108,392	79.65	38	79,506	58.43	31	77,104	56.66
Intergenic regions	7	19,266	14.16	7	49,152	36.12	4	53,635	39.41

NESH (http://www.softberry.com/berry.phtml) trained for human genome annotation (no software trained for bird genome annotation is available). The results are illustrated in Figs. 3C and 3D and listed in Table 2. GENSCAN predicted 52 exons for seven open reading frames and FGENESH predicted 45 exons, also for seven open reading frames. Between the two sets of exons predicted by the two different programs, 35 are identical and 1 is overlapping. We compared the predicted exons with the zebra finch AR cDNA sequence and 5′ RACE sequence and searched them by Blast in the GenBank. Thirty-four exons (31 are predicted by both programs and the other 3 predicted by only GENSCAN) for three genes with cDNA sequence matches in zebra finch or other organisms were confirmed (Fig. 3E and Table 2). The three genes are androgen receptor gene, human oligophrenin 1 homolog, and human hypothetical protein MGC21416 homolog (Fig. 3E from left to right). Table 2 shows that at least 60.59, 3.92, and 56.66% of the BAC sequence encodes genes, exons, and introns, respectively.

The human oligophrenin 1 is a Rho-GTPase activating protein. The human hypothetical protein MGC21416 contains the YIPF6 domain. The YIPF6 protein is a Golgi protein involved in vesicular transport and interacts with GTPases [21]. The human oligophrenin 1 homolog and hypothetical protein MGC21416 homolog might have a functional link. They share a big CpG island (Fig. 3 and Table 1, CpG island 4) at their 5′ ends and might be regulated coordinately. However, no cDNA sequences for these two homologs have been isolated from the zebra finch yet.

We corrected the GENSCAN predicted AR cDNA sequence using the czfAR cDNA sequence and using the BAC nucleotide sequence in a few cases in which it differed from the czfAR sequence. We generated a predicted zebra finch AR cDNA sequence (called pzfAR; GenBank BK005685). The pzfAR cDNA encodes 2445 bp including the poly(A) tail, of which the most 5′ 690 bp are predicted by GENSCAN and the remaining 1755 bp are predicted by GENSCAN and confirmed by the czfAR sequence. Comparison of pzfAR to the BAC sequence indicates that the predicted AR gene is arranged into eight exons. Exon 1 contains 908 bp, of which the most 5′ 690 bp are predicted by GENSCAN and the remaining 218 bp are the most 5′ end of czfAR. Exons 2–8 contain the remainder of czfAR (Table 3). The most 5′ 783 bp of exon 1 are not found in the current draft of chicken chromosome 4, even though chromosome 4 contains 28 other regions (average length 125 bp) of >74% homology to BAC bases 1–33,812, which contain exon 1. We sought to confirm that the 5′ end of exon 1 of pzfAR is expressed in the zebra finch. Using 30 different RT-PCR primer sets (forward primers at the 5′ end of exon 1, reverse primers in czfAR) and several different Taq polymerases and cycle conditions, we attempted to amplify from RNA from three tissues (brain, ovary, and testis). None of these reactions yielded any RT-PCR product, although control RT-PCR within the AR46 region was successful using the same RNA (data not shown). We also designed oligonucleotides encoding 5′ portions of exon 1 and AR46 and used them to probe a Northern blot of RNA from ovary and testis. The AR46 probe recognized the AR in RNA from both tissues, but the 5′ exon 1 probe did not (data not shown). At this time we are unable to confirm that any sequence of pzfAR that is 5′ of czfAR is expressed. Exon 1 is located within a large CpG island (Fig. 3 and Table 1, CpG island 1) at the 5′ end of the predicted AR gene, which may make it difficult to sequence using standard methods or to amplify using RT-PCR. It is noteworthy that the 5′-RACE sequence whose expression was confirmed covers 218 bp of the 3′ end of exon 1. An alignment of the predicted zebra finch androgen receptor shows significant homology to the N-terminal 35 amino acids of mammalian androgen receptors (Fig. 4). However, most of the rest of the 7N-terminal sequence differs from that of other vertebrate

Table 3
Predicted zebra finch androgen receptor exons

pzfARexons	pzfARstart (bp)	Length (bp)	BAC sequence	Homology to czfAR	Homology to chicken Chr. 4	Homology to mouse AR cDNA
1	1	908	33,031–33,938	bp 690–908	bp 783–908	bp 882–908
2	909	152	50,855–51,006	Yes	Yes	Yes
3	1061	117	57,509–57,625	Yes	Yes	Yes
4	1178	288	60,656–60,943	Yes	Yes	Yes
5	1466	145	62,956–63,100	Yes	Yes	Yes
6	1611	131	64,275–64,405	Yes	Yes	Yes
7	1742	158	64,603–64,760	Yes	Yes	Yes
8	1900	524	65,030–65,553	Yes	Not last 274 bp	Not last 368 bp

androgen receptors and is not conserved among the vertebrates (Fig. 4). The most C-terminal approximately 400-amino-acid sequence is, however, closely related to those of other vertebrate androgen receptors.

Table 3 shows the eight exons of the predicted zebra finch AR gene (pzfAR). The first 690 bp are predicted by GENSCAN analysis of the BAC sequence, and the rest of the sequence corresponds to the contig of cDNA sequences (czfAR). The pzfAR 3′ of bp 783 is homologous to the chicken chromosome 4 sequence and 3′ of bp 882 to the mouse AR cDNA. We deleted from the pzfAR two sequences considered to be RACE or cloning artifacts: the first 8 bp of AR46 sequence (GenBank AF532914), which are not found in the RACE product (GenBank AY847476) or BAC 319A15, and the first 27 bp of the RACE product, which are not found in BAC 319A15.

The BAC sequence also allowed the identification of putative estrogen response element (ERE) sequences in and near the AR gene. The ERE defines a locus of interaction of the estrogen receptor with DNA and has the general form GGTCAnnnTGACC. Using Dragon ERE Finder [22], we located a perfect ERE palindrome, GGTCAcccTGACC, at 58,886 bp of the AR BAC 319A15, between AR exons 3 and 4. A similar perfect ERE GGTCAgggTGACC is found between the same two exons of the chicken androgen receptor on chromosome 4 (www.ensemble.org), suggesting

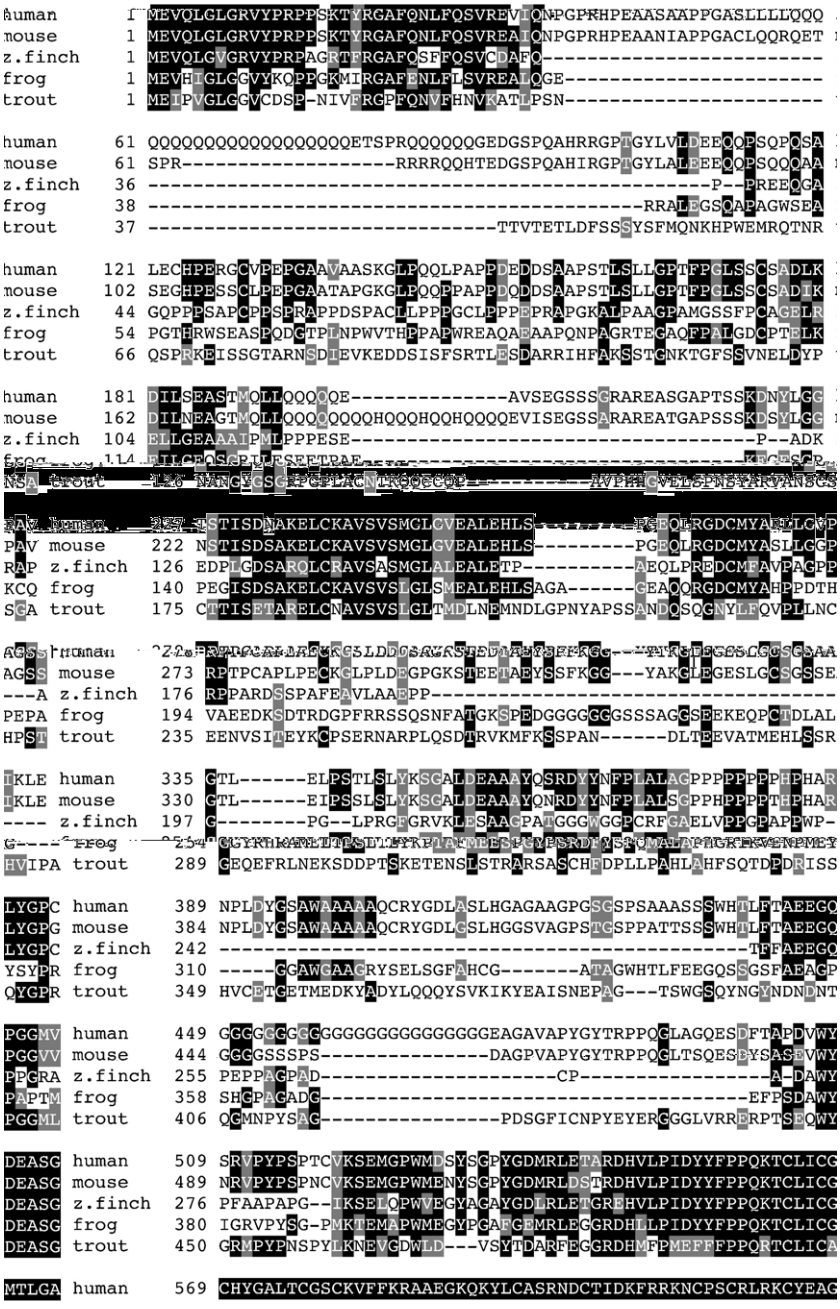

Fig. 4. Comparison of the predicted zebra finch AR with the ARs of human (Homo sapiens, AAA51729), mouse (Mus musculus, AAA37234), African clawed frog (Xenopus laevis, AAC97386), and rainbow trout (Oncorhynchus mykiss, BAA32785).

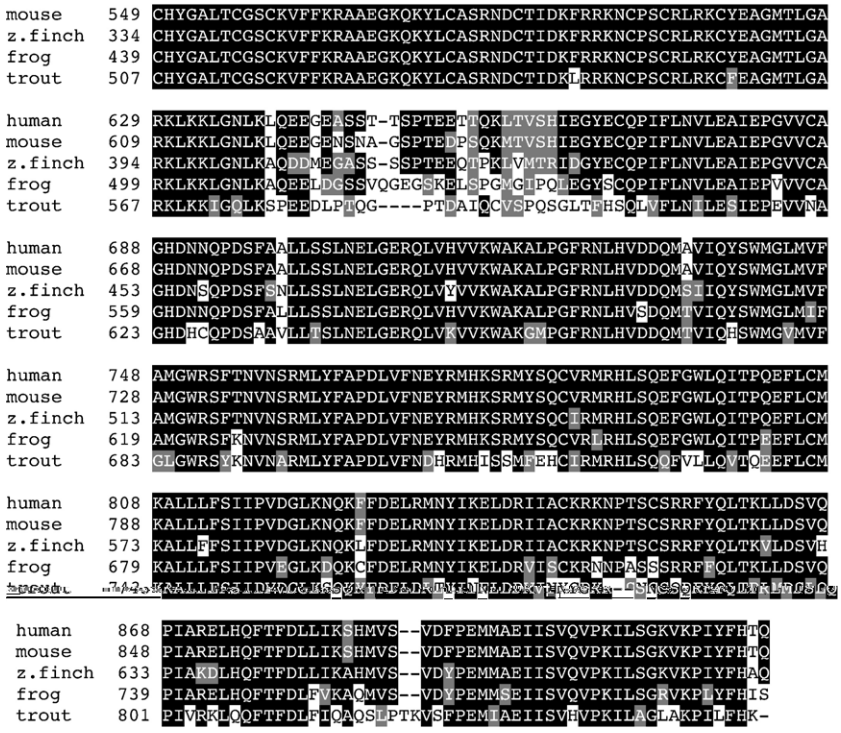

Fig. 4 (continued).

that this sequence has been conserved since the two species diverged about 100 Myr ago [23]. Imperfect ERE palindromes were found 5′ of the pzfAR coding sequence (which starts at 33,031 bp of the BAC sequence), at 28,540 (GGTGAttgTGACT), 13,161 (GGTCAcagCAGCC), and 4548 bp (AGTCAgacTGGCT). These sites are potential sites for estrogen receptor to regulate expression of AR. Estrogens are known to up-regulate or otherwise interact with AR expression in zebra finch brain [24].

The BAC sequence does not contain large tandem duplicates but may contain repeats and transposons (Fig. 5A). Detailed investigation of repeats and transposons is beyond the scope of this paper. When the zebra finch BAC sequence was Blasted and aligned to the chicken genome, it exclusively matched the chicken chromosome 4 approx bp 1,340,000 to 1,200,000 (the complementary strand) throughout the whole range, including nonexonic regions (Fig. 5B). The three genes confirmed in the zebra finch BAC 319A15 were also found in the chicken chromosome 4 1,340,000–1,200,000 bp region with the same relative locations and orientations, indicating again a high synteny between the zebra finch and the chicken genome.

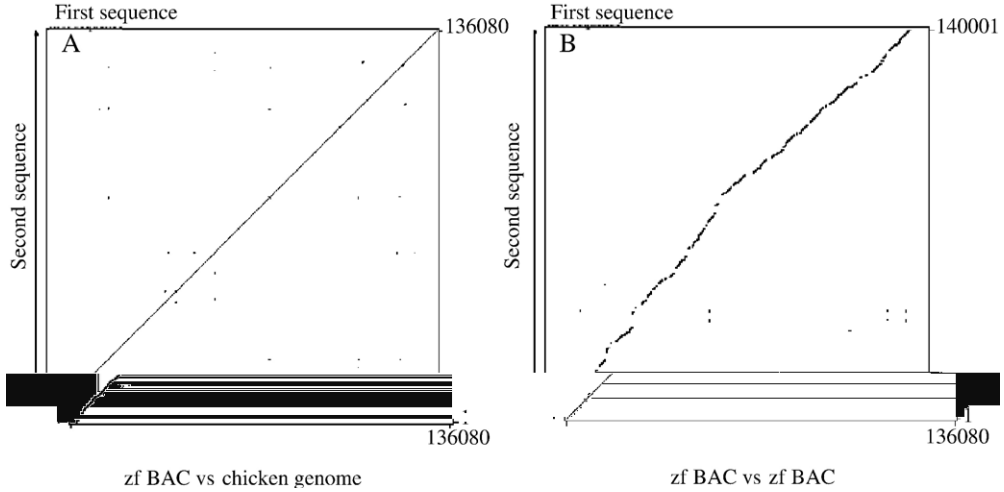

Fig. 5. Dot plot comparison of the zebra finch BAC 319A15 sequence (the first sequence) with (A) itself and (B) the chromosome 4 approx bp 1,340,000 to 1,200,000 of the chicken genome draft sequence (the second sequence). The chicken sequence was retrieved through the Web site http://www.ensembl.org and analyzed by Pipmaker (http://pipmaker.bx.psu.edu/pipmaker/).

Discussion

Arrayed and characterized deep-coverage large-insert BAC libraries are invaluable resources in genomic studies [15–17,25] although nonarrayed BAC libraries are also useful in targeted isolation of genomic regions [26]. We report here the construction of an arrayed high-quality deep-coverage BAC library for zebra finch. The library consists of 147,456 clones, with 98% of them containing inserts of an average size of 134 kb. The haploid genome size of zebra finch is estimated to be 1250 Mb (http://www.genomesize.com) and the library contains 15.5 haploid genome equivalents. The library was deposited in the Arizona Genomics Institute's BAC/EST Resource Center (www.genome.arizona.edu), where the library, filters, and clones are made available to the community on a cost recovery basis.

The high quality and utility of the BAC library were demonstrated by screening of the library filters with 10 single-copy gene probes, BAC end sequencing, and the first prediction of a full-length cDNA encoding an avian androgen receptor. All the 10 probes yielded positive clones, 14 per probe on average. The BAC library has already been used to construct comparative maps of targeted genomic regions using universal overgo probes [27]. Of 102 overgo probes designed from highly conserved sequences between chicken and human, 68 were successful in screening the zebra finch BAC library (see [27] for the criteria). An average of 13.2 zebra finch BAC clones per probe were assigned to the respective contigs by fingerprinting. The numbers of confirmed positive clones per probe were smaller than 15.5-fold haploid genome coverage, possibly because some universal overgo probes (36-bp) might not perfectly match the target sequences and not hybridize optimally.

BAC end sequences are useful for evaluating the quality of BAC libraries [16], developing a sequence-tagged connector (STC) framework [28], surveying genome structures [29], developing SSR [30] and SNP markers [31], constructing comparative genome maps [32], profiling genome rearrangements [33], etc. End sequencing of the positive clones of the Hat3, BDNF, and AR46 probes showed that about 94% of BAC ends produced high-quality sequences with an average high-quality base pair number of 533 bp. If all clones of the whole BAC library are sequenced at both ends at the same success rate and high-quality base pair number, it will provide an average of one sequence (STC) per 4.3 kb and 11.6% of coverage of the zebra finch genome. Comparison of the BAC end sequences with the chicken genome draft sequence indicated that the zebra finch genome is highly syntenic to the chicken genome. An average of 62.2% BAC end sequences matched to the corresponding regions in the chicken genome draft sequence. The failure of match of about 37.8% BAC end sequences could have several explanations. Some sequences might be of low quality; some might be derived from nonconserved regions or insertion regions of the zebra finch genome, and some might fall in regions of the chicken genome draft sequence that are not sequenced or have low sequence quality. Also, we cannot exclude the possibility that some BAC clones were chimeras. Neither end sequence of the AR BAC 319A15 (Fig. 1) matched to chicken genome draft sequence. This clone had been chosen for whole BAC sequencing before knowing the BAC end sequence analysis results. However, the 136-kb BAC sequence matched to the corresponding region in the chicken genome draft sequence throughout the whole range (Fig. 5). It is worth noting that the two end sequences (~1 kb) account for only about 0.7% of the whole BAC sequence.

BAC libraries have been used as critical tools in physical mapping [4,5], genome sequencing [6,7], and positional cloning [8,9]. Here we provide a typical example that a BAC library is also an important tool to identify and clone genes that are difficult to characterize by other methods. Despite the high interest in the avian androgen receptor, a full-length AR cDNA has not been isolated from any bird species, although several laboratories have used a variety of techniques to isolate partial cDNAs [18,19]. The sequencing here of the AR-containing BAC has yielded important new information concerning the avian androgen receptor.

Our analysis of the single BAC sequence has uncovered a previously unknown 35-amino-acid sequence at the putative amino terminus of the AR protein, which has high homology to mammalian androgen receptors. Previous studies have measured the expression of androgen receptor proteins in numerous avian species, using several different antibodies, which are all directed against portions of the amino-terminal 40 amino acids of the mammalian androgen receptor [34–38]. These antibodies recognize avian ARs in zebra finches and other passerine species, and in quail. The utility of these antibodies suggests that the amino-terminal sequence of the zebra finch AR, which we have predicted from the BAC genomic sequence by GENSCAN, is likely correct, even though we were unable to confirm the expression of this sequence in zebra finches using oligonucleotide probes and RT-PCR. Although the corresponding nucleotide sequence has not yet been found in the androgen receptor gene on chicken chromosome 4, we suspect that the absence of the sequence may be due to incomplete sequence information for this region. The same antibodies recognizing this region of the protein are also successful for immunolocalization of the androgen receptor in chickens [36,38]. The predicted sequence of the zebra finch androgen receptor reported here, however, implies strongly that after the first 35 amino acids, the zebra finch and mammalian androgen receptors share little homology except in a region corresponding to the C-terminal half of the mammalian androgen receptor, which contains the highly conserved DNA and ligand binding domains [39]. Further work is needed to confirm the accuracy and expression of our predicted 5′-end sequence of the first exon of the AR gene. Exon 1 is about 17 kb distant from exons 2–8 and lies within a large CpG island.

The predicted androgen receptor gene is contained fully within a single BAC. The human oligophrenin 1 homolog and human hypothetical protein MGC21416 homolog share a promoter region and may have a functional link. They are also contained in a single BAC. BACs containing a full gene or functionally linked genes are powerful tools to study gene functions by, for example, genetic transformation [40]. Modified genes are usually used in genetic transformation, and techniques for DNA manipulation on BACs have been established [41].

Comparative genome analysis among species with optimal phylogenetic distances is a powerful approach to explore genome structure and functions [42]. The differences among species must be reflected in differences among genomes. Our sequence analysis of BAC ends and a single BAC illustrates the utility of the BAC library, because we identified an avian full-length androgen receptor orf and a conserved regulatory element (ERE) in the AR gene, and found that the zebra finch genome is highly syntenic to the chicken genome (Figs. 2 and 5). Because of the high homology of the genomic sequence of these two species, and the availability of the chicken genome sequence as a reference, it will be possible to use BAC fingerprinting [43,44] and end sequencing to construct a physical map of the zebra finch genome, an invaluable tool for functional and evolutionary studies of songbirds and other species.

Materials and methods

BAC library construction

The BAC library was constructed with HindIII using a method adapted from Osoegawa and de Jong [25] and Luo and Wing [15]. Genomic DNA was prepared from livers of several individual female zebra finches. Fresh livers were homogenized in ice-cold 1× PBS buffer (0.75% NaCl, 0.02% KCl, 0.144% Na_2HPO_4, 0.024% KH_2PO_4, pH 7.0) in a Dounce tissue grinder (Wheaton). After passing through one layer of Miracloth (Calbiochem) to remove tissue chunks, cells were collected by centrifugation at ~200g (1000 rpm for Beckman GS-6R centrifuge) at 4°C, adjusted to ~5×10^8 cells/ml with 1× PBS, and embedded at a 1:1 (v/v) ratio with 1.0% low-melting temperature agarose prepared in 1× PBS (final concentration of 0.5%) in 80-μl plug molds (Bio-Rad). The linearized, dephosphorylated single-copy BAC vector was prepared from the high-copy pCUGIBAC1 as described by Luo et al. [45]. Other procedures of the BAC library construction exactly followed our protocols [15]. DH10B T1-resistant cells (Invitrogen) were used for transformation.

BAC library screening

The whole BAC library was gridded onto eight 22.5 × 22.5-cm filters in high-density, double spots and 4 × 4 patterns with Genetix Q-Bot (Genetix). Each 22.5 × 22.5-cm filter supports 18,432 clones in duplicate in six fields. Ten single-copy gene-specific probes were used to screen the BAC library filters. They were zRalDH (AF162770), BDNF (AF255389), NR2A (AB042757), NR2B (AB107125), FoxP2(AY549148), ZENK (AF026084), HAT3(L33860), ZF1A (S75898), TrkB (AY679520), and AR46(AF532914). Protocols for high-density BAC library filter screening and address determination of positive signals are publicly available from our Web site (www.genome.arizona.edu). Sometimes a mixture of probes was used in screening. Putative positive clones from the library screening were verified with individual probes by colony or Southern hybridization using standard techniques [46].

BAC end sequencing

BAC DNA was isolated from 1.2 ml 2× YT (Fisher) overnight culture with Tomtec Quadra 96 Model 320 (Tomtec) in a 96-well format. Isolated BAC DNA was sequenced at both ends using BigDye Terminator v.3 (Applied Biosystems) according to the manufacturer's instruction. The T7 primer (5′-TAATACGACTCACTATAGGG-3′) was used as the "forward" primer and the BES_HR primer (5′-CACTCATTAGGCACCCCA-3′) was used as the "reverse" primer. Cycle sequencing was performed using PTC-200 thermal cyclers (MJ Research) in a 384-well format with the following regime: 150 cycles of 10 s at 95°C, 5 s at 55°C, and 2.5 min at 60°C. After the cycle-sequencing step, the DNA was purified by magnetic beads, CleanSeq (http://www.agencourt.com), according to the manufacturer's instruction. Samples were eluted into 20 μl of water and separated on ABI 3730xl DNA capillary sequencers with default conditions. Sequence data were collected by data collection software (Applied Biosystems), extracted using sequence analysis software (Applied Biosystems), and transferred to a UNIX workstation. Sequences were base-called using the program Phred [47,48]; vector and low-quality (Phred value <16) sequences were removed by CROSS_MATCH [47,48]. The BAC end sequences were mapped to the chicken genome draft sequence through the Web site http://www.ensembl.org/Multi/blastview?species=Gallus_gallus.

Shotgun sequencing, finishing, and sequence analysis

The zebra finch BAC 319A15 plasmid was isolated using a modified alkaline lysis method (http://www.genome.arizona.edu/information/protocols/BAC_DNA_prep.html) and randomly sheared by using a hydroshear device (Gene Machine). End repair was performed by using a DNA end repair kit (Epicentre) according to the manufacturer's directions. DNA fragments 2–4 kb in size were size-fractionated and ligated into the EcoRV site of pBluescript II KS(+) vector (Stratagene). Ligated DNA was transformed into Escherichia coli DH10B cells via electroporation, and recombinant clones were randomly picked using Q-Bot (Genetix). DNA templates for sequencing were generated from 150 μl 2× YT (Fisher) overnight culture with Tomtec Quadra 96 Model 320 (Tomtec) in a 96-well format and sequenced at both ends using BigDye Terminator v.3 (Applied Biosystems) according to the manufacturer's instruction. T7 and T3 were used as primers. Cycle sequencing was performed using PTC-200 thermal cyclers (MJ Research) in a 384-well format with the following regime: 35 cycles of 10 s at 96°C, 5 s at 50°C, and 4 min at 60°C. Procedures from PCR product purification to base calling are the same as for BAC end sequencing. Sequences were assembled using PHRAP (http://bozeman.genome.washington.edu/phrap.docs/phrap.html). The software CONSED [49] was used to view the assembly and finish the sequence. Shotgun clones spanning gaps were picked and completely sequenced by bacterial transposon-mediated method [50] to join contigs. The primer walking method was also applied to extend contigs and improve sequence quality and the software CONSED was used to design the primers. Some GC-rich regions were sequenced with the Sequencing Finishing Kit from Amersham. A finished contig with the quality of Phred 40 or greater was finally checked by comparing in silico digestion patterns with real digestion patterns of NotI, HindIII, BamHI, XhoI, SwaI, SalI, and PvuI. The BAC sequence was analyzed online through www.ebi.ac.uk/emboss/cpgplot, http://genes.mit.edu/GENSCAN.html, http://www.softberry.com/berry.phtml, and http://pipmaker.bx.psu.edu/pipmaker. The chicken genome draft sequence was retrieved through the Web site http://www.ensembl.org/Multi/blastview?species=Gallus_gallus.

Acknowledgments

We thank Jetty S.S. Ammiraju, Olin Feuerbacher, Samina Makda, Angelina Angelova, and Teri Rambo of the Arizona Genomics Institute for technical assistance and Kiran Rao of AGCOL for database management. Thanks to David Clayton (University of Illinois), Erich Jarvis (Duke University Medical Center), and Claudio Mello (Oregon Health & Science University) for assistance, encouragement, and cDNA probes. This work was supported by grants from the NIH (Grants U1HG02525A and DC000217).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.09.005.

References

[1] D.F. Clayton, Songbird genomics: methods, mechanisms, opportunities, and pitfalls, Ann. N. Y. Acad. Sci. 1016 (2004) 45–60.
[2] F. Nottebohm, T.M. Stokes, C.M. Leonard, Central control of song in the canary (Serinus canarius), J. Comp. Neurol. 165 (1976) 457–486.
[3] H.P. Ziegler, P. Marler, Behavioral Neurobiology of Birdsong, N.Y. Acad. Sci, New York, 2004.

[4] J.D. McPherson, et al., A physical map of the human genome, Nature 409 (2001) 934–941.
[5] J.W. Wallis, et al., A physical map of the chicken genome, Nature 432 (2004) 761–764.
[6] E.S. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[7] R.H. Waterston, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[8] D. Botstein, N. Risch, Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease, Nat. Genet. 33 (2003) 228–237.
[9] J.S. Bodnar, et al., Positional cloning of the combined hyperlipidemia gene Hyplip1, Nat. Genet. 30 (2002) 110–116.
[10] J.S. Gasper, T. Shiina, H. Inoko, S.V. Edwards, Songbird genomics: analysis of 45 kb upstream of a polymorphic Mhc class II gene in red-winged blackbirds (Agelaius phoeniceus), Genomics 75 (2001) 26–34.
[11] L.W. Hillier, et al., Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution, Nature 432 (2004) 695–716.
[12] Y. Itoh, A.P. Arnold, Chromosomal polymorphism and comparative painting analysis in the zebra finch, Chromosome Res. 13 (2005) 47–56.
[13] H. Shizuya, et al., Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector, Proc. Natl. Acad. Sci. USA 89 (1992) 8794–8797.
[14] J.P. Tomkins, et al., New genomic resources for the honey bee (Apis mellifera L.): development of a deep-coverage BAC library and a preliminary STC database, Genet. Mol. Res. 1 (2002) 306–316.
[15] M. Luo, R.A. Wing, An improved method for plant BAC library construction, in: E. Grotewold (Ed.), Plant Functional Genomics: Methods and Protocols, Humana Press, Totowa, NJ, 2003, pp. 3–19.
[16] K. Osoegawa, et al., BAC resources for the rat genome project, Genome Res. 14 (2004) 780–785.
[17] T. Miyake, C.T. Amemiya, BAC libraries and comparative genomics of aquatic chordate species, Comp. Biochem. Physiol. C Toxicol. Pharmacol. 138 (2004) 233–244.
[18] K.L. Nastiuk, D.F. Clayton, The canary androgen receptor mRNA is localized in the song control nuclei of the brain and is rapidly regulated by testosterone, J. Neurobiol. 26 (1995) 213–224.
[19] W.R. Perlman, B. Ramachandran, A.P. Arnold, Expression of androgen receptor mRNA in the late embryonic and early posthatch zebra finch brain, J. Comp. Neurol. 455 (2003) 513–530.
[20] H.A. McQueen, et al., CpG islands of chicken are concentrated on microchromosomes, Nat. Genet. 12 (1996) 321–324.
[21] A. Marchler-Bauer, et al., CDD: a conserved domain database for protein classification, Nucleic Acids Res. 33 (2005) D192–D196.
[22] V.B. Bajic, et al., Dragon ERE Finder version 2: a tool for accurate detection and analysis of estrogen response elements in vertebrate genomes, Nucleic Acids Res. 31 (2003) 3605–3607.
[23] M. van Tuinen, S.B. Hedges, Calibration of avian molecular clocks, Mol. Biol. Evol. 8 (2001) 206–213.
[24] Y.H. Kim, W.R. Perlman, A.P. Arnold, Expression of androgen receptor mRNA in the zebra finch song system: developmental regulation by estrogen, J. Comp. Neurol. 469 (2004) 535–547.
[25] K. Osoegawa, P.J. de Jong, BAC library construction, in: S. Zhao, M. Stodolsky (Eds.), Bacterial Artificial Chromosomes, Library Construction, Physical Mapping, and Sequencing, vol. 1, Humana Press, Totowa, NJ, 2004, pp. 1–46.
[26] E. Isidore, et al., Direct targeting and rapid isolation of BAC clones spanning a defined chromosome region, Funct. Integr. Genom. 5 (2005) 97–103.
[27] W.A. Kellner, R.T. Sullivan, B.H. Carlson, J.W. Thomas, Uprobe: a genome-wide universal probe resource for comparative physical mapping in vertebrates, Genome Res. 15 (2005) 166–173.
[28] G.G. Mahairas, et al., Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome, Proc. Natl. Acad. Sci. USA 96 (1999) 9739–9744.
[29] L.F. Marek, et al., Soybean genomic survey: BAC-end sequences near RFLP and SSR markers, Genome 44 (2001) 572–581.
[30] X. Qi, et al., An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, Pennisetum glaucum, Theor. Appl. Genet. 109 (2004) 1485–1493.
[31] M.M. Weil, R. Pershad, R. Wang, S. Zhao, Use of BAC end sequences for SNP discovery, in: S. Zhao, M. Stodolsky (Eds.), Bacterial Artificial Chromosomes, Functional Studies, vol. 2, Humana Press, Totowa, NJ, 2004, pp. 1–6.
[32] R.A. Wing, et al., The Oryza Map Alignment Project: the golden path to unlocking the genetic potential of wild rice species, Plant Mol. Biol. 59 (2005) 53–62.
[33] S. Volik, et al., End-sequence profiling: sequence-based analysis of aberrant genomes, Proc. Natl. Acad. Sci. USA 100 (2003) 7696–7701.
[34] J. Balthazart, A. Foidart, M. Houbart, G.S. Prins, G.F. Ball, Distribution of androgen receptor-immunoreactive cells in the quail forebrain and their relationship with aromatase immunoreactivity, J. Neurobiol. 35 (1998) 323–340.
[35] G.T. Smith, E.A. Brenowitz, G.S. Prins, Use of PG-21 immunocytochemistry to detect androgen receptors in the songbird brain, J. Histochem. Cytochem. 44 (1996) 1075–1080.
[36] B.A. Shanbhag, P.J. Sharp, Immunocytochemical localization of androgen receptor in the comb, uropygial gland, testis, and epididymis in the domestic chicken, Gen. Comp. Endocrinol. 101 (1996) 76–82.
[37] K.K. Soma, V.N. Hartman, J.C. Wingfield, E.A. Brenowitz, Seasonal changes in androgen receptor immunoreactivity in the song nucleus HVc of a wild bird, J. Comp. Neurol. 409 (1999) 224–236.
[38] B.K. Shaw, G.G. Kennedy, Evidence for species differences in the pattern of androgen receptor distribution in relation to species differences in an androgen-dependent behavior, J. Neurobiol. 52 (2002) 203–220.
[39] H.E. MacLean, G.L. Warne, J.D. Zajac, Localization of functional domains in the androgen receptor, J. Steroid Biochem. Mol. Biol. 62 (1997) 233–242.
[40] M. Kamihira, K. Nishijima, S. Iijima, Transgenic birds for the production of recombinant proteins, Adv. Biochem. Eng. Biotechnol. 91 (2004) 171–189.
[41] S. Gong, X.W. Yang, C. Li, N. Heintz, Highly efficient modification of bacterial artificial chromosomes (BACs) using novel shuttle vectors containing the R6Kgamma origin of replication, Genome Res. 12 (2002) 1992–1998.
[42] E.H. Margulies, et al., Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes, Proc. Natl. Acad. Sci. USA 102 (2005) 3354–3359.
[43] C. Soderlund, S. Humphray, A. Dunham, L. French, Contigs built with fingerprints, markers, and FPC V4.7, Genome Res. 10 (2000) 1772–1787.
[44] M.C. Luo, et al., High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis, Genomics 82 (2003) 378–389.
[45] M. Luo, et al., Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon Fusarium wilt (Fom-2), Genome 44 (2001) 154–162.
[46] J. Sambrook, D.W. Russell (Eds.), Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001.
[47] B. Ewing, P. Green, Base-calling of automated sequencer traces using Phred: II. Error probabilities, Genome Res. 8 (1998) 186–194.
[48] B. Ewing, L. Hillier, M.C. Wendl, P. Green, Base-calling of automated sequencer traces using Phred: I. Accuracy assessment, Genome Res. 8 (1998) 175–185.
[49] D. Gordon, C. Abajian, P. Green, Consed: a graphical tool for sequence finishing, Genome Res. 8 (1998) 195–202.
[50] S. Haapa, et al., An efficient DNA sequencing strategy based on the bacteriophage mu in vitro DNA transposition reaction, Genome Res. 9 (1999) 308–315.