The Plant Genome
Activities & Resources

Construction of an *Amaranthus hypochondriacus* Bacterial Artificial Chromosome Library and Genomic Sequencing of Herbicide Target Genes

Peter J. Maughan,* Nicholas Sisneros, Meizhong Luo, Dave Kudrna, Jetty S. S. Ammiraju, and Rod A. Wing

P.J. Maughan, Dep. of Plant and Wildlife Sciences, Brigham Young Univ., Provo, UT, 84602; N. Sisneros, M. Luo, D. Kudrna, J.S.S. Ammiraju, and R.A. Wing, Arizona Genomics Institute, Dep. of Plant Sciences, Univ. of Arizona, Tucson, AZ 85721. Received 8 Aug. 2007. *Corresponding author (Jeff_Maughan@byu.edu).

Abstract

Before the Spanish conquest of the ancient Americas, the grain amaranths (*Amaranthus caudatus* L., *A. cruentus* L., *A. hypochondriacus* L.) were a staple food of the New World. Recently, the grain amaranths have regained international attention for their nutritional quality and importance as a symbol of indigenous cultures. Here we report the development of a bacterial artificial chromosome (BAC) library constructed from the cultivar 'Plainsman' (*A. hypochondriacus*; $2n = 2x = 32$). The library consists of a total of 36,864 clones with an average insert size of 147 kb with less than 1.8% of the clones containing empty vectors. The frequency of BAC clones carrying inserts derived from chloroplast and mitochondrial DNA was estimated to be 6.9%. Thus, based on a haploid genome size of 466 Mb per haploid nucleus, the BAC library coverage is approximately 10.6 times the haploid genome content. The genome coverage estimate was empirically confirmed by screening the library with seven low copy amaranth probe sequences. The utility of the amaranth BAC library was demonstrated by identification and full-length genomic sequencing of the acetolactate synthase and protoporphyrinogen oxidase genes—both major targets for several classes of important herbicides. The quality of the BAC library for BAC end sequencing projects was evaluated by bidirectional end sequencing of 384 random clones. End sequences were annotated using BLAST searches and queries to plant transposable element databases.

Published in Crop Sci. 48(S1) S85–S94. Published 8 Feb. 2008.
doi:10.2135/cropsci2007.08.0434tpg

677 S. Segoe Rd., Madison, WI 53711 USA

THE GENUS *AMARANTHUS* (Caryophyllales: Amaranthaceae) encompasses about 60 species with worldwide distribution (Sauer 1976). Three species of the genus produce edible seeds. These grain amaranths (*A. hypochondriacus* L., *A. cruentus* L., and *A. caudatus* L.) were important food staples among the ancient civilizations of Central and South America. They remain important as a food crop in several areas of modern Latin America and have recently received substantial attention as an alternative food crop (Bressani et al., 1992). Among the favorable characteristics of the grain amaranths are the quality and quantity of protein in the seed. Amaranth seed protein is exceptionally high in lysine (generally the first limiting amino acid in the cereal grains) with an amino acid composition that compares favorably with FAO/WHO standards (Becker et al., 1981). Amaranth seed protein content ranges from 12.5 to 22.5% on a dry matter basis, with an average of about 15%, a value that is notably higher than cereal grains (Bressani, 1989; Breene, 1991).

The genus also includes several well-known weedy species, collectively referred to as "pigweeds," which

Abbreviations: ALS, acetolactate synthase; BAC, bacterial artificial chromosome; BES, BAC end sequence; CHEF, contour-clamped homogeneous electric field; HMW, high molecular weight; PCR, polymerase chain reaction; PPO, protoporphyrinogen oxidase; SSR, simple sequence repeat; UTR, untranslated region.

are arguably the most damaging weeds in the United States. Two pigweeds, redroot pigweed (*A. retroflexus* L.) and slender amaranth (*A. viridis* L.), are among the most widely distributed weeds in the world (Holm et al., 1997), while smooth pigweed (*A. hybridus* L.) and spiny amaranth (*A. spinosus* L.) are ranked among the 18 most serious weeds in the world (Holm et al., 1991). Epitomizing their weed status, the pigweeds are notorious for their ability to develop resistance to herbicides. Biotypes resistant to triazine and acetolactate synthase (ALS)-inhibiting herbicides have been reported for the majority of the major pigweed species present in the United States (Heap, 1997). In addition, one or more weedy amaranth species has been reported with resistance to dinitroanilines (e.g., trifluralin [α,α,α-trifluoro-2,6-dinitro-*N,N*-dipropyl-*p*-toluidine]), bipyridyliums (e.g., paraquat [1,1′-dimethyl-4,4′-bipyridinium]), and protoporphyrinogen oxidase (PPO) inhibitors {e.g., acifluorfen [5-(2-chloro-α,α,α-trifluoro-*p*-tolyloxy)-2-nitrobenzoic acid]} (Heap, 2004).

Notwithstanding the importance of the grain amaranths as emerging alternative crop species and the relative importance of the related weedy species, only a few molecular investigations have been reported for the Amaranthaceae and even fewer molecular tools, needed for advanced genomic studies, have been developed. Genomic analysis tools, such as sequence-based molecular markers (e.g., microsatellites), genetic linkage maps, and expressed sequence tagged collections have yet to be developed for any of the *Amaranthus* species. One tool of particular importance for future genomic studies in amaranth is a bacterial artificial chromosome (BAC) library. Bacterial artificial chromosome libraries are critical for identifying full length genomic sequences and correlating genetic and physical maps, and for comparative genomics, and are the vital first step toward whole genome sequencing projects (Monaco and Larin, 1994). Bacterial artificial chromosome libraries have been successfully developed for numerous plant species, including economically important crop species, secondary or emerging crop species, and model organisms (Woo et al., 1994; Yoo et al., 2003; Tomkins et al., 2001, 2004; Wang et al., 2005; Stevens et al., 2006). Here we describe the construction of the first BAC library for the genus *Amaranthus*. In characterizing the library we (i) determined the average insert size and organellar DNA content of the library, (ii) calculated the haploid equivalence of the library using low-copy genomic probes, (iii) evaluated the utility of the library for gene discovery by screening the library for two herbicide genes targeted (ALS and PPX2L), and (iv) obtained a glimpse of the amaranth genome by bidirectionally sequencing 384 random BAC clones.

Materials and Methods

Plant Material

Seeds for the *A. hypochondriacus* cultivar Plainsman were kindly provided by D. Baltensperger (Texas A&M University, College Station, TX). Plants were grown at 25°C with a 12-h photoperiod, in Sunshine Mix II (Sun Grow, Inc., Bellevue, WA), supplemented weekly with N fertilizer. All plants were grown in 15-cm pots in greenhouses at Brigham Young University, Provo, UT.

BAC Library Construction

Fifty grams of young leaf material from 30-d-old plants were flash frozen in liquid N and stored at −80°C before nuclei preparation. Purified nuclei and partial *Hin*dIII restriction enzyme digestion (0.4 U for 20 min at 37°C) of the high molecular weight (HMW) DNA were performed as described by Luo and Wing (2003). The partially digested HMW DNA was size-fractionated twice by separation in 1% agarose contour-clamped homogeneous electric field (CHEF) gels (Bio-Rad Laboratories, Hercules, CA) with a 1- to 50-s linear ramp and 6 V cm^{-1}, at 14°C in 0.5× TBE for 20 h to obtain HMW DNA fragments ranging in size from 130 to 400 kb. The library was constructed by ligating 180 ng of HMW DNA with 20 ng dephosphorylated *Hin*dIII cloning-ready pAGIBAC1 vector at 16°C in an overnight reaction as described by Luo et al. (2001). ElectroMax DH10B T$_1$ phage resistant *E. coli* cells (Invitrogen, Carlsbad, CA) were transformed with ligation mix via electroporation as described by Luo and Wing (2003). Cells were plated on selective LB medium containing 12.5 μg mL^{-1} of chloramphenicol, 80 μg mL^{-1} X-gal, and 100 μg mL^{-1} isopropyl β-D-1-thiogalactopyranoside and incubated overnight at 37°C. Transformed colonies were picked and transferred robotically to barcode-ordered 384-well microtiter plates, grown at 37°C overnight, and stored at −80°C. The master library and one copy are stored at the Resource Center of the Arizona Genomics Institute, while a third copy of the library is stored at Brigham Young University. The library, named AH_PBa, is available to the public for a nominal processing fee (http://genome.arizona.edu/orders/ verified 21 Jan. 2008).

BAC Clone Insert Analysis

To determine the insert size distribution and average insert size of the library, BAC DNA from 384 randomly selected clones were isolated using Tomtec

Quadra 96 model 320 (Tomtec, Hamden, CT) using a modified alkaline lysis method. The samples were then digested with the restriction enzyme *Not*I and electrophoretically separated on 1% agarose CHEF gels as described above. The insert sizes for each clone was determined by comparison with molecular weight standards.

BAC End Sequencing

Bacterial artificial chromosome DNA for the same 384 clones used for insert size characterization were sequenced bidirectionally (768 sequences) using the BES_HR primer in the reverse direction (CAC TCA TTA GGC ACC CCA) and the standard T7 in the forward direction with BigDye Terminator v.3.1 cycle sequencing on ABI 3730 xl DNA Analyzers (Applied Biosystems, ABI, Foster City, CA). Base-calling and vector sequence removal were performed using the computer programs Phred (Ewing and Green, 1998) and CROSS-MATCH (Ewing et al., 1998). High quality BAC end sequences (BESs), defined as sequences with at least 100 bases of Phred quality scores >20, were annotated using BLASTN, BLASTX, and queries to plant repeat element databases using RepeatMasker 3.0 (http://www.repeatmasker.org; verified 21 Jan. 2008).

BAC Library Screening

The library was double-spotted onto two high-density Hybond N+ membranes (Amersham, Piscataway, NJ) with a Q-bot (Genetix, New Milton, UK) robot in a four by four pattern according to the method described by Luo et al. (2006). Each membrane contains 18,432 clones in duplicate in six fields. Hybridization probes were labeled with α-dCTP^{32} using a Prime-a-Gene kit (Promega, Madison, WI) and membranes were hybridized using standard protocols as described by Sambrook et al. (1989). Chloroplast DNA content was determined by screening the library with a mixture of three barley (*Hordeum vulgare* L.) chloroplast gene sequences (*ndhA*, *rbcL*, and *psbA*), while the mitochondrial DNA content was determined by screening the library with a mixture of four mitochondrial gene sequences (*coxE*, *cob*, *atpA*, *atp9*) as described by Ammiraju et al. (2006).

Genome equivalency was also analyzed by screening the library with seven low-copy nuclear gene sequences identified from the BES analysis (Table 1). Primers designed from the BES were used to amplify portions of each of these gene sequence from genomic DNA using HotStar *Taq* Master Mix polymerase chain reaction (PCR) kit according to manufacture's protocol (Qiagen, Valencia, CA). Each gene probe was labeled as previously described and hybridized to both the BAC library and a genomic DNA Southern blot. For Southern blot analysis, genomic DNA was extracted from *A. hypochondriacus* (Plainsman) plants using the protocol described by Todd and Vodkin (1996). Six micrograms of genomic DNA was digested separately with one of three restriction enzymes (*Eco*RI, *Eco*RV, *Hin*dIII), size fractioned, and transferred to Hybond-N+ positively charged nylon membrane (Amersham, Piscataway, NJ) using a neutral transfer protocol as described by the manufacture. All filters were hybridized at 65°C and washed under stringent conditions at 65°C in 0.1× saline sodium citrate, 0.1% sodium dodecyl sulfate.

ALS and PPO Genes

Polymerase chain reaction primer sequences for amplifying the ALS and PPO (PPX2L) genes from *A. hypochondriacus* (Plainsman) were kindly provide by P. Tranel (University of Illinois, Champaign-Urbana, IL; Table 1). The primer sequences were based on cDNA sequences for the ALS gene from *A. retroflexus*, *A. powellii* S. Wats. (AAK50820, AAK50821; McNaughton et al., 2005; Diebold et al., 2003), and *A. tuberculatus* (Moq.) Sauer (ABD52329; Patzoldt et al., 2006) for the PPX2L gene. Amplified PCR products were labeled and used to screen the BAC library as described above. Bacterial artificial chromosome plasmid from a single positive clone for each gene was extracted using a NucleoBond BAC 100 kit (Macherey-Nagel, Easton, PA). Genomic sequence of the entire ALS and PPX2L genes were obtained via primer walking techniques on the respective BAC plasmid. DNA sequencing for the primer walking was performed at the Brigham Young University DNA sequence center (Provo, UT) using standard ABI Prism *Taq* dye-terminator cycle sequencing methodology. DNA sequence chromatograms were analyzed with the Contig Express program in the Vector NTI software suite (InforMax, Frederick, MD).

Results and Discussion

BAC Library Construction and Characterization

We constructed a BAC library from *A. hypochondriacus* (Plainsman) which consisted of 36,864 clones ordered into 96 384-well microtiter plates. The library was developed by partially digesting HMW DNA with the endonuclease enzyme *Hin*dIII, followed by CHEF gel electrophoresis size selection for fragments in the 130- to 220-kb range as described by Ammiraju et al. (2006) and Luo and Wing (2003). This gel fragment was then subjected to a second round of CHEF gel electrophoresis from which two gel fragments, an

A1 and an A2 fragment, were isolated with DNA sizes ranging from 120 to 175 kb and 175 to 220 kb, respectively. High molecular weight DNA from both the A1 and A2 gel fragments were electro-eluted, ligated into the pAGIBACI cloning vector, and transformed into DH10B *E. coli* cells. Based on a random initial screening of 24 BAC clones derived from each of the two gel fractions, we estimated the average insert size of the A1 fraction to be 128 kb, while that of the A2 fraction 157 kb (Fig. 1a). The ligation and transformation efficiency of the larger A2 fragments was much lower than that of the smaller A1 fragments and necessitated the picking of colonies from both fractions to complete the full BAC library. Thus, 17,664 BAC clones were picked from the A2 fraction, while remaining 19,200 BAC clones were picked from the A1 fraction.

To estimate the average insert size of the final BAC library, 384 BAC clones (four clones from each of the 96 microtiter plates in the library) were chosen at random for BAC plasmid isolation, *Not*I digestion, and CHEF gel separation (Fig. 1b). Of the 384 BAC clones analyzed, two yielded no DNA and seven (1.8%) lacked an insert (empty vector). Of the remaining 375 BAC clones, the average insert size of the clones with inserts was 147 ± 41 kb (SD) with the insert sizes ranging from 12 to 354 kb. Ninety-three percent of the clones have an insert sizes above 100 kb. The distribution of insert sizes is shown in Fig. 2. Internal *Not*I recognition sites within the cloned amaranth DNA, identified as two or more insert bands on the CHEF gels, were identified in 14% of the BAC clones (Fig. 1) and is consistent with the low percentage of *Not*I recognition sites (GCGGCCGC) found in BAC libraries of other dicotyledonous species (Choi et al., 1995; Noir et al., 2004; Tomkins et al., 1999; Hong et al., 2004). In contrast, Stevens et al. (2006) reported that BAC clones from quinoa (*Chenopodium quinoa* Willd.), the closest relative

Table 1. Results of screening the amaranth bacterial artificial chromosome (BAC) library. Probes 1 to 7 were derived from BAC end sequences that showed high sequence homology to known nuclear genes. Probes 8 and 9 were derived from primers designed to putative conserved regions of the acetolactate synthase and protoporphyrinogen oxidase (PPX2L) genes.

	Probe name	Primer sequences (5′→3′)	Expected product size (bp)	GenBank accession no.[†] (E-value)	Probe function or putative function	No. of bands detected by Southern hybridization[‡]	No. of positive signals
1	AH-PBa0005_B19	(for) TCGGGCTGTTATGACTTTCC (rev) AGGAGGCAACGTAACCATTG	648	AAM98245.1 (1.4E-47)	Pentaxin	2	23
2	AH-PBa0020_B19	(for) CTAAAGCACCGAGCAGAGGT (rev) CTGGGCTAAAGAAGCCAGTG	389	AAM65192.1 (2.5E-20)	Diphosphomevalonate decarboxylase	3	27
3	AH-PBa0037_B21	(for) ACCACGTGCAATTTCCCTAA (rev) AGAGGCAACATCCGACAATC	511	ABE41799.1 (2.2E-13)	Tocopherol cyclase	2	27
4	AH-PBa0015_B19	(for) TTCCCTCCACAATGCTTACC (rev) ACACCGTAGCGGAGTGTACC	761	NP_568967.1 (5.3E-26)	Pectate lyase family protein	2	46
5	AH-PBa0054_B21	(for) ATCCAGATTTGGCTGACAGG (rev) TACATGGGCTGCATGATGTT	442	NP_179734.1 (4.0E-102)	Coatomer protein complex, subunit alpha	1	6
6	AH-PBa0066_B17	(for) TGTCACGGTGCAGAAAGAAG (rev) CCAAAGTCACCAACATGTGC	586	CAB61983.1 (7.0E-57)	Receptor-kinase like protein	2	2
7	AH-PBa0075_B17	(for) GAGATATCGGCCCAGACAAA (rev) CGTTGGGATTTGGAGAAGAA	451	NM_117537.2 (7.0E-22)	Acylaminoacyl-peptidase related protein	2	43
					Subtotal average	2	24.0
					Genome equivalents	12.4×	
8	ALS	(for) AGCTCTTGAACGTGAAGGTG (rev) TCAATCAAAACAGGTCCAGG	444	NA	Acetolactate synthase	5	46
9	PPX2L	(for) AGAAAACGCAATGCTACTGAG (rev) ACAACCTCCAGAAAATTTTG	712	NA	Protoporphyrinogen oxidase	3	42
					Average	2.4	30
					Genome equivalents	12.3×	
10	Chloroplast probe set	*ndhA, rbcL, psbA*					2705
11	Mitochondrial probe set	*coxE, cob, atpA, atp9*					36
					Sum		2741
					Organellar contamination		6.9%

[†] Based on GenBank queries using BLASTX limited to the *Viridiplantae* (taxid: 33090).

[‡] The number of bands detected by Southern hybridization of genomic DNA digested with *Eco*RI, *Eco*RV, or *Hin*dIII. The number of bands reported was the maximum number of detected bands in one of the digests.

of *A. hypochondriacus* with a reported BAC library, showed a significantly higher prevalence (38–44%) of internal *Not*I recognition sites which they attributed to the presence of a highly repetitive DNA element containing the *Not*I recognition motif. The lack of *Not*I recognition sites in *A. hypochondriacus*, however, suggests that if indeed a highly repetitive DNA element is responsible for the increased number of *Not*I recognition sites in *C. quinoa*, that DNA element is not present in high abundance in the amaranth genome or the amplification of the element in the Amaranthaceae occurred after the evolutionary divergence of the chenopods and the amaranths.

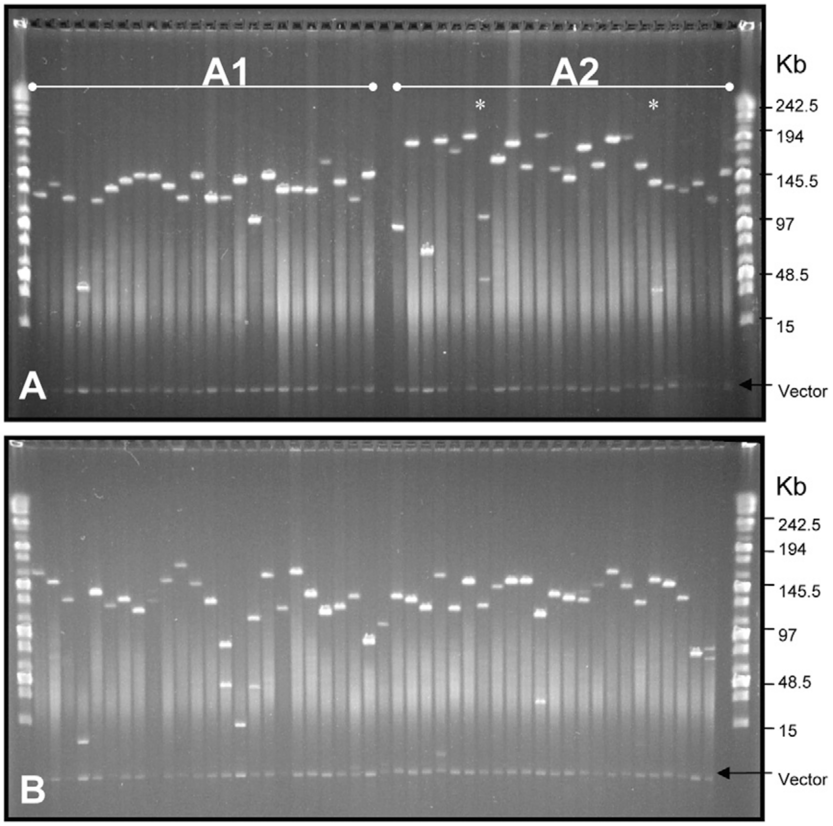

Figure 1. *Not*I digestion of bacterial artificial chromosome (BAC) DNA isolated from an amaranth library. (A) An ethidium bromide–stained contour-clamped homogeneous electric field (CHEF) gel of 24 random BAC clones from the A1 (left) and A2 (right) gel fragments. Average insert size of the A1-derived fragments was 128 kb, while the average insert size of the A2-derived fragments was 157 kb. BAC clones with internal *Not*I recognition sites are identified with an asterisk. (B) An ethidium bromide–stained CHEF gel of 48 random BAC clones from the final amaranth BAC library. The first and last lanes in both figures contain size markers, given in kilobases.

Organellar Content and Size Estimation

To estimate the level of contamination of the library with organellar DNA and to further our ability to identify specific DNA sequences in the library we robotically prepared double-spotted high density DNA arrays of the library on nylon membranes. To calculate the level of organellar DNA contamination in the library, we probed the membranes with a series of chloroplast and mitochondrial genes (Table 1). Based on the hybridization signals, we calculate the chloroplast DNA contamination of the library to be 6.8% (2507 clones) and mitochondrial DNA to be 0.1% (36 clones). Based on an organellar contamination of 6.9% and an estimate of 1.8% nonrecombinants (no inserts; see above), we calculate that the BAC library contains approximately 33,657 clones that are derived from nuclear DNA and 2543 that are derived from organellar DNA (mostly cpDNA). Spinach (*Spinacia oleracea* L.; AJ400848), the closest relative of *A. hypochondriacus* with a completely sequenced plastid genome, has a chloroplast genome size of 151 kb (Schmitz-Linneweber et al., 2001). Thus with an average BAC insert size of the A2 fraction being 157 kb, it is possible that some of the large cpBAC clones (e.g., clone AH-Pba0-009-b23 estimated to be 151 kb) in the library may contain entire chloroplast genomes and may prove to be particularly useful for studying the amaranth chloroplast genome.

Since *A. hypochondriacus* has a diploid genome DNA content of 0.95 pg, which corresponds to a haploid genome size of 466 Mb per haploid nucleus (Bennett and Smith, 1991), the predicted nuclear genome coverage of the BAC library is 10.6× genome equivalents. Thus, there is greater than a 99% probability that any specific *A. hypochondriacus* DNA sequence is present in the library (Liu, 1998). To empirically test the library for coverage we screened the library with seven probes developed from BESs that showed high sequence homology (E-value > 1.0E-10; BLASTX searches against the nonredundant GenBank database) to known nuclear genes (Table 1). Probes for the hybridization analysis were PCR amplified from genomic DNA. Each of these PCR amplifications produced a single, strong amplification product (as visualized on agarose gels) and Southern

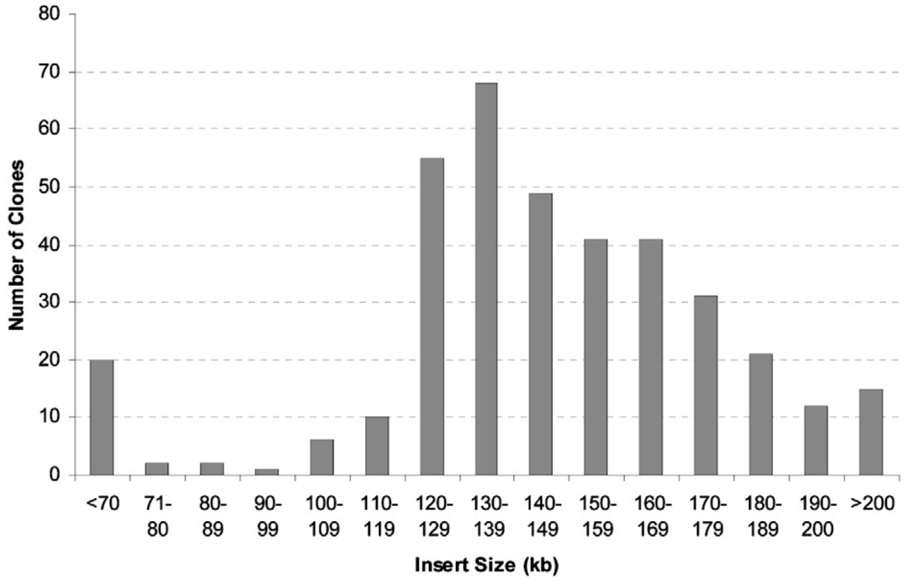

Figure 2. Insert size distribution of 374 randomly selected bacterial artificial chromosome (BAC) clones.

blot hybridization to genomic DNA confirmed that these sequences are present in one to three copies in the *A. hypochondriacus* genome (Table 1; probes 1–7). The average number of bands detected in the Southern blots was two, while the average number of BAC clones that hybridized to these probes was 24.9, suggesting that the genome equivalents represented by the BAC library is ~12.4×. While this estimate is slightly higher than that predicted based on average insert size and number of BAC clones, it is probably slightly biased upward since the probes used for the analysis are derived from BAC clones known to be in the library. In all cases the BAC clone from which the probe sequence was derived showed a positive hybridization signal. The discrepancy may also be due to an overestimation of the genome size (466 Mb per haploid nucleus) or it may reflect the presence of duplicated loci for some of the genes in the amaranth genome, consistent with its paleo-allotetraploid evolution (Pal and Khoshoo, 1974). Thus, we believe that 10.6× genome equivalents for the BAC library is a conservative estimate.

BAC End Sequencing

To examine the quality of the library for sequencing and to get a glimpse of the *A. hypochondriacus* genome, we bidirectionally end sequenced 384 BAC clones which produced 728 reads with an average length of 747 high quality bases (Phred quality >20) calls per sample for a total of 563 kb of sequence data. The BESs have been submitted to GenBank under accession numbers ER914893 to ER915620. The G+C content was 35%. Sequence data screened with the computer program RepeatMasker v.3.1.6 using the *Arabidopsis* repeat database revealed 15,998 bp (2.84%) of the sequence was retroelements, while a much smaller amount (408 bp) were DNA transposons. The vast majority of the retroelements were classified as Ty1/Copia-like LTR elements, while hobo-Activator elements were the most common DNA transposon (Table 2). The fraction of predicted repeat sequences identified is low for a complex genome and suggests that *Amaranthus* may possess unique repeated sequences not detected by comparisons to the *Arabidopsis* repeat database.

RepeatMasker also identified 65 simple sequence repeats (SSRs), which included 7 mono-, 15 di-, 23 tri-, and 10 tetra-nucleotide repeats with the dominant nucleotide repeat motif in each class being the $(A)_n$, $(AT)_n$, $(AAT)_n$, and $(ATTT)_n$ motifs, respectively (Table 2). The average number of repeat unit (n) within each SSR locus was ~17, but varied from a low of three for a TGGGGG hexa-nucleotide repeat (ER915595) to a high of 177 for a TA dinucleotide repeat (ER914984). Both perfect and imperfect SSR motif patterns were identified. Currently no sequence-based genetic markers exist for *A. hypochondriacus*, thus the discovery of SSRs are of particular interest. Simple sequence repeats are highly polymorphic, exhibit codominant inheritance, and are simple to assay using PCR (Morgante et al., 2002). Simple sequence repeats have been the marker system of choice for developing genetic linkage maps and should prove particularly useful elucidating the evolutionary origins of the grain amaranths from their putative progenitor species (*A. hybridus*, *A. quitensis* H.B.K., and *A. powellii*). The observed abundance of AT-rich repeat motifs is consistent with previous reports of motif abundance in other dicotyledonous plant species, including other species of the Amaranthaceae family (*Beta vulgaris* L., Mörchen et al., 1996; spinach, Groben and Wricke, 1998; *C. quinoa*, Mason et al., 2005) and suggests that future efforts at large scale development of SSR markers from enriched libraries should focus on AT-based motifs (e.g., AT and TAA).

BLASTN searches against the *Arabidopsis* chloroplast genome (GenBank ID AP000423) using the high-quality BESs, masked for low complexity and repeat sequences, identified 97 sequences derived from 49 BAC clones as likely chloroplast-derived (E-value > 1.0E-6). In all cases where a BES showed significant homology to the *Arabidopsis* chloroplast genome, its paired BES also showed significant homology to the chloroplast genome, suggesting that these clones were not nuclear introgressions of chloroplast DNA. Based on these BLAST results, the level of chloroplast-derived sequences in the library may actually be as high as 12.8%, suggesting that if the library is used for high throughput applications (e.g., fingerprinting for physical map construction), a round of re-arraying for the removal of redundant chloroplast clones may be economically beneficial.

To search for coding sequences within our BES collection, we screened 631 nonredundant BESs without homology to the chloroplast genome against the *Arabidopsis* RefSeq protein database. Of these, 173 (27%) were homologous to at least one protein sequence (E-value > 1.0E-6), of which 40 (19%) were annotated as "unknown" or "hypothetical" proteins. The remaining 458 BESs were then screened against the RefSeq and the nonredundant (*nr*) protein database for all green plants (*Viridiplantae*) where an additional 41 (6.5%) BESs had detectable homologs to at least one protein sequence (E-value > 1.0E-6). Thus a total of 214 (34%) of the BES had detectable homologs in the RefSeq protein database. Based on an estimated genome size of 466 Mb per haploid nucleus for *A. hypochondriacus*, and on the average *Arabidopsis* gene length of 2 kb (Arabidopsis Genome Initiative, 2000), a cautious estimation of the total cDNA coding capacity for *A. hypochondriacus* is 158 Mb and a total gene content is 79,021 (Lai et al., 2006). We note that this estimate is tentative, as it assumes several key variables, including accurate prediction of genome size and average gene length, as well as the successful identification of all homologs via BLASTX searches of GenBank.

Table 2. Summary of the plant repeat element content of 728 high-quality (563,867 bp) amaranth bacterial artificial chromosome (BAC) end sequences as determined using the RepeatMasker (V. 3.1.8) and the default RepDB for *Arabidopsis*.

Class	Element[†]	No. of elements[‡]	Length (bp) and % of sequence
Retroelements	SINEs	0	15,998 (2.84%)
	LINEs		
	L1/CIN4	5	
	LTR elements		
	Ty1/Copia	25	
	Gypsy/DIRS1	7	
DNA transposons	Hobo-Activator	3	408 (0.07%)
	En-Spm	1	
	MuDR-IS905	1	
	Total interspersed repeats	42	16,406 (2.91%)
Simple sequence repeats			
Mono-	$(T)_n$	6	
	Other[§]	1	
Di-	$(AT)_n$	14	
	Other[§]	1	
Tri-	$(TAA)_n$	12	
	$(CAT)_n$	2	
	$(GAA)_n$	2	
	$(TGG)_n$	3	
	Other[§]	4	
Tetra-	$(TAAA)_n$	6	
	$(TTAA)_n$	3	
	Other[§]	1	
Penta-	$(TTTTA)_n$	2	
	Other[§]	5	
Others	NA	4	
	Total number of SSRs	66	4296 (0.76%)

[†] SINE, short interspersed nuclear elements; LINE, long interspersed nuclear elements; LTR, long terminal repeat; NA, not applicable; SSRs, simple sequence repeats.

[‡] Most repeats fragmented by insertions or deletions were counted as one element.

[§] Each SSR motifs in the other category was found a single time.

ALS and PPO Genes

An interesting feature of the *Amaranthus* genus is that it includes several well-known weedy species which are arguably the most damaging weeds in the United States. Many of these weedy *Amaranthus* species have developed genetic resistance to important herbicides, including those that target ALS and PPO genes. Resistant and susceptible classes of cDNA sequences for the ALS gene have been reported from *A. retroflexus* and *A. powellii* (McNaughton et al., 2005; Diebold et al., 2003), while cDNA sequences for the PPO gene have been reported for *A. tuberculatus* (Patzoldt et al., 2006). While all previous work analyzed cDNA sequence, we used the *A. hypochondriacus* BAC library to identify genomic sequence of these herbicide target enzymes, including 5′ untranslated region (UTR), coding, intron, and 3′ UTR portions. Using primers derived from conserved regions of the cDNA sequences (Table 1) from these weedy species we amplified homologs of ALS and PPO (e.g., PPX2L isoform) genes from *A. hypochondriacus* and used these homologs to screen the BAC library. From this hybridization, we identified 46 and 42 positive BAC clones for the ALS and PPX2L genes, respectively. Southern hybridizations using these same probes against genomic DNA from *A. hypochondriacus* identified five bands for the ALS probe and three bands for the PPO probe, suggesting that multiple copies of these sequences reside within the *A. hypochondriacus* genome (Table 1).

To sequence the ALS and PPX2L genes, a positive BAC clone for each gene was selected and primer walking was used to sequence directly off the BAC DNA. On the ALS BAC clone, we sequenced a total of 3299 bp, which included 561, 2010, and 728 bp of 5′ UTR, coding/intron, and 3′ UTR DNA sequences, respectively. Using the cDNA sequences from *A. retroflexus* (AF363369), *A. powellii* (AF363370), and *A. tuberculatus* (ABM53018) as references, the *A. hypochondriacus* ALS gene was predicted to be intronless and translate into a protein of 669 amino acids with a molecular weight of 72.87 kDa. Comparing the *A. hypochondriacus* ALS coding sequence with the ALS sequences of these three weedy *Amaranthus* species showed that the ALS gene is highly conserved, with 72 nucleotide changes which altered only 19 amino acids, 11 of which are in the chloroplast transit peptide region (Table 3). McNaughton et al. (2005) identified five amino acid changes that characterized the change from ALS susceptible to resistant in *A. retroflexus* and *A. powellii*. The *A. hypochondriacus* sequence contains all five of these amino acids in the susceptible state of the protein and would presumably be susceptible to ALS-targeted herbicides. The genomic sequence for the ALS gene has been submitted to GenBank (EU024568).

Protoporphyrinogen oxidase converts protoporphyrinogen IX to protoporphyrin IX and is the last common enzyme in the tetrapyrrole biosynthetic pathway that produces heme and chlorophyll (Beale and Weinstein, 1990). Three different nuclear PPO genes are known to function in *A. tuberculatus* (Patzoldt et al., 2006). The PPX1 gene codes for a plastid-targeted PPO isozyme, the PPX2 gene encodes a mitochondria-targeted PPO isozymes, while the PPX2L gene encodes both plastid- and mitochondria-targeted PPO isoforms due to the presence of an alternate in-frame initiation codon (Patzoldt et al., 2006). Resistance to the PPO-targeted herbicides has only been documented in the PPX2L gene (Patzoldt et al., 2006). Using a PPX2L probe, we identified a PPX2L-containing BAC clone and sequenced the PPX2L gene from *A. hypochondriacus*. From this clone we sequenced 16,391 bp, which included 2233, 11,716, and 2442 bp of 5′ UTR, coding/intron, and 3′ UTR DNA, respectively. The complete genomic sequence for the PPX2L gene has been submitted to GenBank (EU024569). Using the cDNA sequence for PPX2L from *A. tuberculatus* (DQ386117) we aligned our genomic DNA and predicted the presence of 18 exons comprising 1608 bp of coding sequence (including the stop codon) and 10,108 bp of intron sequence. The average exon size was 89 bp and ranged from 37 to 174 bp, while the average intron size was 595 bp and ranged from 68 to 2696 bp. When translated, the coding sequence generated a protein of 535 amino acids with a molecular weight of 58.75 kDa. In agreement with the observations of Patzoldt et al. (2006) and Watanabe et al. (2001) we observed a putative alternative start site (Met31) at amino acid position 31. Characteristic of chloroplast targeting sequences, the N-terminal extension sequence (first 30 amino acids) did not contain any acidic residues (Asp or Glu) but was enriched with Ser and Thr. The computer program ChloroP predicted the chloroplast targeting sequences was contained within the first 30 AA sequence (Emanuelsson et al., 1999). Comparison of the *A. hypochondriacus* PPX2L coding sequence with susceptible (DQ386117) and resistant (DQ386116) isoforms of the PPX2L gene from *A. tuberculatus* identified 55 nucleotide changes which altered 16 amino acids (Table 4). Included in these nucleotide differences is a 3-bp deletion (ΔG210) in the ninth exon which deletes a glycine residue in resistant *A. tuberculatus* biotypes and is believed to be responsible for PPO-targeted herbicide resistance. As expected, the PPX2L gene from *A. hypochondriacus* does not have the ΔG210 mutation and is susceptible to PPO-targeted herbicides.

Table 3. Summary of sequence variation in the acetolactate synthase (ALS) gene among four *Amaranthus* species. Italicized amino acid positions lie within the chloroplast transit peptide region. Amino acid deletions are designated with a dash. Sequences were aligned using ClustalW.

GenBank accession no.†	Species	Amino acid position																		
		4	*6*	*7*	*11*	*12*	*17*	*29*	*38*	*44*	*84*	*89*	174	270	321	346	419	423	479	622
AAK50820	*A. retroflexus*	T	S	N	S	S	N	I	T	L	S	E	L	Y	E	H	Q	K	V	A
AAK50821	*A. powellii*	T	S	N	S	S	N	I	T	L	S	E	L	Y	E	H	Q	K	V	T
ABM53018	*A. tuberculatus*	T	–	Q	S	S	N	L	A	L	F	D	F	F	K	Q	R	N	I	A
EU024568	*A. hypochondriacus*	N	S	N	F	Y	Y	I	T	P	S	E	L	Y	E	H	Q	K	V	A
	Consensus	T	S	N	S	S	N	I	T	L	S	E	L	Y	E	H	Q	K	V	A

†All accession sequences are from ALS-herbicide susceptible biotypes.

Table 4. Summary of sequence variation in the protoporphyrinogen oxidase gene (PPX2L) sequence between *A. hypochondriacus* ('Plainsman') and resistant and susceptible biotypes of *A. tuberculatus*. Sequences were aligned using ClustalW. Amino acid deletions are designated with a dash.

GenBank accession no.	Species	Amino acid position															
		12	68	133	148	170	210	218	224	228	253	270	284	432	447	526	531
ABD52329[†]	*A. tuberculatus*	N	S	A	T	H	G	M	E	I	–	Q	L	N	S	I	D
ABD52328[‡]	*A. tuberculatus*	N	S	D	T	H	–[§]	M	E	I	–	Q	L	N	C	I	D
EU024569	*A. hypochondriacus*	K	N	D	S	R	G	V	D	V	G	H	I	K	S	L	N

[†] Susceptible biotype.

[‡] Resistant biotype.

[§] The deletion of a single glycine residue at position 211 confers resistance to herbicides inhibiting protoporphyrinogen oxidase (Patzoldt et al., 2006).

Conclusions

To our knowledge this is the first BAC library reported for the genus *Amaranthus* (Caryophllales: Amaranthaceae). The genus encompasses over 60 species with worldwide distribution and includes grain crops, ornamentals, and many damaging weedy species. The BAC library presented here contains approximately 10 genome equivalents and has an average insert size of 147 kb. The utility of the library for identifying full-length genomic clones for genes of interest from across the genus was demonstrated by identifying, sequencing, and characterizing two herbicide-targeted genes (ALS and PPX2L) that have been overcome in the weedy *Amaranthus* species. The next goal of the project is to use the BAC library to develop a physical map of the amaranth genome. The development of a physical map should help us elucidate the genetic basis for several traits of particular interest within the genus, including C_4 photosynthesis, monoecious and dioecious sex differentiation, and extreme abiotic stress tolerance.

Acknowledgments

This research was funded by grants from the Ezra Taft Benson Agriculture and Food Institute and the Holmes Family Foundation. We gratefully acknowledge D. Baltensperger (Texas A&M University) for seed contribution and P. Tranel (University of Illinois) for providing primers for the ALS and PPX2L genes.

References

Ammiraju, J.S.S., M. Luo, J.L. Goicoechea, W. Wang, D. Kudrna, C. Mueller, J. Talag, H. Kim, N.B. Sisneros, B. Blackmon, E. Fang, J.B. Tomkins, D. Brar, D. MacKill, S. McCouch, N. Kurata, G. Lambert, D.W. Galbraith, K. Arumuganathan, K. Rao, J.G. Walling, N. Gill, Y. Yu, P. SanMiguel, C. Soderlund, S. Jackson, and R.A. Wing. 2006. The *Oryza* bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*. Genome Res. 16:140–147.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

Beale, S.I., and J.D. Weinstein. 1990. Tetrapyrrole metabolism in photosynthetic organisms. p. 287–391. *In* H.A. Dailey (ed.) Biosynthesis of heme and chlorophylls. McGraw-Hill, New York.

Becker, R., E.L. Wheeler, K. Lorenz, A.E. Stafford, O.K. Grosjean, A.A. Betchart, and R.M. Saunders. 1981. A compositional study of amaranth grain. J. Food Sci. 46:1175–1180.

Bennett, M.D., and J.B. Smith. 1991. Nuclear DNA amounts in angiosperms. Philos. Trans. R. Soc. London, Ser. B. 334:309–345.

Breene, W.M. 1991. Food uses of grain amaranth. Cereal Foods World 36:426–430.

Bressani, R. 1989. The proteins of grain amaranth. Food Rev. Int. 5:13–38.

Bressani, R., A. Sanchez-Marroquin, and E. Morales. 1992. Chemical composition of grain amaranth cultivars and effects of processing on their nutritional quality. Food Rev. Int. 8:23–49.

Choi, S., R.A. Creelman, J.E. Mullet, and R.A. Wing. 1995. Construction and characterization of a bacterial artificial chromosome library of *Arabidopsis thaliana*. Plant Mol. Biol. Rep. 13:124–128.

Diebold, R.S., K.E. McNaughton, E.A. Lee, and F.J. Tardif. 2003. Multiple resistance to imazethapyr and atrazine in Powell amaranth (*Amaranthus powellii*). Weed Sci. 51:312–318.

Emanuelsson, O., H. Nielsen, and G. von Heijne. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. 8:978–984.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using Phred: II. Error probabilities. Genome Res. 8:186–194.

Ewing, B., L. Hillier, M. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. 8:175–185.

Groben, R., and G. Wricke. 1998. Occurrence of microsatellites in spinach sequences form computer databases and development of polymorphic SSR markers. Plant Breed. 117:271–274.

Heap, I.M. 1997. The occurrence of herbicide-resistant weeds worldwide. Pestic. Sci. 51:235–243.

Heap, I.M. 2004. International survey of herbicide resistant weeds. Available at http://www.weedscience.org (verified 16 Jan. 2008).

Holm, L., J. Doll, E. Holm, J. Pancho, and J.P. Herberger. 1997. World weeds: Natural histories and distribution. John Wiley & Sons, New York.

Holm, L.G., D.L. Plucknett, J.V. Pancho, and J.P. Herberger. 1991. The world's worst weeds: Distribution and biology. Krieger, Malabar, FL.

Hong, C.P., S.J. Lee, J.Y. Park, P. Plaha, Y.S. Park, Y.K. Lee, J.E. Choi, K.Y. Kim, J.H. Lee, J. Lee, H. Jin, S.R. Choi, and Y.P. Lim. 2004. Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences. Mol. Genet. Genomics 271:709–716.

Lai, C.W.J., Y. Qingyi, S. Hou, R.L. Skelton, M.R. Jones, K.L.T. Lewis, J. Murray, M. Eustice, P. Guan, R. Agbayani, P.H. Moore, R. Ming, and G.G. Presting. 2006. Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome. Mol. Genet. Genomics 276:1–12.

Liu, B.H. 1998. Statistical genomics: Linkage, mapping, and QTL analysis. CRC Press, Boca Raton, FL.

Luo, M., H. Kim, D. Kudrna, N.B. Sisneros, S.-J. Lee, C. Mueller, K. Collura, A. Zuccolo, E.B. Buckingham, S.M. Grim, K. Yanagiya, H. Inoko, T. Shiina, M.F. Flajnik, R.A. Wing, and Y. Ohta. 2006. Construction of a nurse shark (*Ginglymostoma cirratum*) bacterial artificial chromosome (BAC) library and a preliminary genome survey. BMC Genomics 7:106.

Luo, M., Y.-H. Wang, D. Frisch, T. Joobeur, R.A. Wing, and R.A. Dean. 2001. 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: 154–162.

Luo, M., and R. Wing. 2003. An improved method for plant BAC library construction. Methods Mol. Biol. 236:3–20.

Mason, S.L., M.R. Stevens, E.N. Jellen, A. Bonifacio, D.J. Fairbanks, C.E. Coleman, R. McCarty, A. Rassmussen, and P.J. Maughan. 2005. Development and use of microsatellite markers for germplasm characterization in *Chenopodium quinoa* Willd. Crop Sci. 45:1618–1630.

McNaughton, K.E., J. Letarte, E.A. Lee, and F.J. Tardif. 2005. Mutations in ALS confer herbicide resistance in redroot pigweed (*Amaranthus retroflexus*) and Powell amaranth (*Amaranthus powellii*). Weed Sci. 53:17–22.

Monaco, A.P., and Z. Larin. 1994. YACs, BACs, PACs, and MACs: Artificial chromosomes as research tools. Trends Biotechnol. 12:280–286.

Mörchen, M., J. Cuguen, G. Michaelis, C. Hänni, and P. Saumitoue-Laprade. 1996. Abundance and length polymorphism of microsatellite repeats in *Beta vulgaris* L. Theor. Appl. Genet. 92:326–333.

Morgante, M., M. Hanafey, and W. Powell. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. 30:194–200.

Noir, S., S. Patheyron, M.C. Combes, P. Lashermes, and B. Chalhoub. 2004. Construction and characterisation of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.). Theor. Appl. Genet. 109:225–230.

Pal, M., and T.N. Khoshoo. 1974. Grain amaranths. p. 129–137. *In* J.B. Hutchinson (ed.) Evolutionary studies in world crops: Diversity and change in the Indian subcontinent. Cambridge Univ. Press, Cambridge, UK.

Patzoldt, W.L., A.G. Hager, J.S. McCormick, and P.J. Tranel. 2006. A codon deletion confers resistance to herbicides inhibiting protoporphyrinogen oxidase. Proc. Natl. Acad. Sci. USA 103:12329–12334.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. Molecular cloning: A laboratory manual. 2nd ed. Cold Spring Harbor Laboratory Press, Plainview, NY.

Sauer, J.D. 1976. Grain amaranths. p. 4–7. *In* N.W. Simmonds (ed.) Evolution of crop plants. Longman, London.

Schmitz-Linneweber, C., R.M. Maier, J.P. Alcaraz, A. Cottet, R.G. Herrmann, and R. Mache. 2001. The plastid chromosome of spinach (*Spinacia oleracea*): Complete nucleotide sequence and gene organization. Plant Mol. Biol. 45:307–315.

Stevens, M.R., C.E. Coleman, S.E. Parkinson, P.J. Maughan, H.-B. Zhang, M.R. Balzotti, D.L. Kooyman, K. Arumuganathan, A. Bonifacio, D.J. Fairbanks, E.N. Jellen, and J.J. Stevens. 2006. Construction of a quinoa (*Chenopodium quinoa* Willd.) BAC library and its use in identifying genes encoding seed storage proteins. Theor. Appl. Genet. 112:1593–1600.

Todd, J.J., and L.O. Vodkin. 1996. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. Plant Cell 8:687–699.

Tomkins, J.P., M. Fregene, D. Main, H. Kim, R. Wing, and J. Tohme. 2004. Bacterial artificial chromosome (BAC) library resource for positional cloning of pest and disease resistance genes in cassava (*Manihot esculenta* Crantz). Plant Mol. Biol. 56:555–561.

Tomkins, J.P., R. Mahalingam, H. Smith, J.L. Goicoechea, H.T. Knap, and R.A. Wing. 1999. A bacterial artificial chromosome library for soybean PI 437654 and identification of clones associated with cyst nematode resistance. Plant Mol. Biol. 41:25–32.

Tomkins, J.P., D.G. Peterson, T.J. Yang, D. Main, E.F. Ablett, R.J. Henry, L.S. Lee, T.A. Holton, D. Waters, and R.A. Wing. 2001. Grape (*Vitis vinifera* L.) BAC library construction, preliminary STC analysis, and identification of clones associated with flavonoid and stilbene biosynthesis. Am. J. Enol. Vitic. 52:287–291.

Wang, W., M. Tanurdzic, M. Luo, N. Sisneros, H.R. Kim, J.-K. Weng, D. Kudrna, C. Mueller, K. Arumuganathan, J. Carlson, C. Chapple, C. de Pamphilis, D. Mandoli, J. Tomkins, R.A. Wing, and J.A. Banks. 2005. Construction of a bacterial artificial chromosome library for spikemoss *Selaginella moellendorffii*: A new resource for plant comparative genomics. BMC Plant Biol. 5:10.

Watanabe, N., F.-S. Che, M. Iwano, S. Takayama, S. Yoshida, and A. Isogai. 2001. Dual targeting of spinach protoporphyrinogen oxidase II to mitochondria and chloroplasts by alternative use of two in-frame initiation codons. J. Biol. Chem. 276:20474–20481.

Woo, S.S., J. Jiang, B.S. Gill, A.H. Paterson, and R.A. Wing. 1994. Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res. 22:4922–4931.

Yoo, E.Y., S. Kim, Y.H. Kim, C.J. Lee, and B.D. Kim. 2003. Construction of a deep coverage BAC library from *Capsicum annuum*, ‘CM334’. Theor. Appl. Genet. 107:540–543.