2015-06-01

The Amaranth (Amaranthus Hypochondriacus) Genome: Genome, Transcriptome and Physical Map Assembly

Jared William Clouse
Brigham Young University - Provo

The Amaranth (*Amaranthus hypochondriacus*) Genome:

Genome, Transcriptome and Physical Map Assembly

Jared William Clouse

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

P. Jeffery Maughan, Chair
Eric N. Jellen
Joshua A. Udall

Department of Plant and Wildlife Sciences

Brigham Young University

June 2015

ABSTRACT

The Amaranth (*Amaranthus Hypochondriacus*) Genome:
Genome, Transcriptome and Physical Map Assembly

Jared William Clouse
Department of Plant and Wildlife Sciences, BYU
Master of Science

Amaranthus hypochondriacus is an emerging pseudo-cereal native to the New World which has garnered increased attention in recent years due to its nutritional quality, in particular its seed protein, and more specifically its high levels of the essential amino acid lysine. It belongs to the Amaranthaceae family, is an ancient paleotetraploid that shows amphidiploid inheritance (2n=32), and has an estimated genome size of 466 Mb. Here we present a high-quality draft genome sequence of the grain amaranth *A. hypochondriacus.* The genome assembly consisted of 377 Mb in 3,518 scaffolds with an N_{50} of 371 kb. Repetitive element analysis predicted that 48% of the genome is comprised of repeat sequences, of which *Copia*-like elements were the most common classified retrotransposon. A transcriptome, consisting of 66,370 contigs, was assembled from eight different tissue and abiotic stress libraries. Annotation of the genome identified 23,059 genes that were supported by our *de novo* transcriptome assembly, the RefBeet 1.1 gene index and the Uniprot_sprot database. To describe the genetic diversity within the grain amaranths (*A. hypochondriacus, A. caudatus,* and *A. cruentus*) and their putative progenitor (*A. hybridus*) we re-sequenced seven accessions in the genus *Amaranthus* (four *A. hypochondriacus*, and one of each *A. caudatus, A. cruentus,* and *A. hybridus*), which identified 7,184,636 and 1,760,433 interspecific and intraspecific single nucleotide polymorphisms, respectively. A phylogeny analysis of the re-sequenced accessions substantiated the classification of *A. hybridus* as the progenitor species of the grain amaranths. Lastly, we generated a physical map for *A. hypochondriacus* using the BioNano optical mapping platform. The physical map spanned 340 Mb and a hybrid assembly using the BioNano optical genome maps nearly doubled the N_{50} of the assembly to 697 kb. Moreover, we analyzed synteny between amaranth and *Beta vulgaris* (sugar beet) and estimated, using Ks analysis, the age of the most recent polyploidization event in amaranth.

Key words: amaranth, physical map, transcriptome, whole-genome sequencing, re-sequencing

ACKNOWLEDGMENTS

First and most importantly, I would like to express my gratitude to my family and especially my wonderful wife for her patience, love, and support throughout the duration of this project. I would also like to express my appreciation to my advisor Dr. Jeff Maughan for allowing me to undertake this project and for his guidance, support, and sharing of his expertise with me throughout my time at BYU both as an undergraduate and graduate student. I also thank the other members of my graduate committee Dr. Josh Udall and Dr. Rick Jellen for their assistance and willingness to participate in this expedited project. Thanks to Justin Page for tutoring and advising me as I learned bioinformatics. This project would not have been possible without his bioinformatic expertise. Thanks also to our collaborators Michael Deyholos, Dinesh Adhikary, and Thiru Ramaraj. Thanks to all the undergraduate students who assisted on this project and to my fellow graduate students for their kindness, friendship, and always lively conversation in our shared office space.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: The Amaranth (*Amaranthus Hypochondriacus*) Genome: Genome, Transcriptome and Physical Map Assembly

INTRODUCTION

The grain amaranths (*Amaranthus hypochondriacus* L.*, Amaranthus caudatus* L.*,* and *Amaranthus cruentus* L.) are part of the Amaranthaceae family (Caryophyllales: Amaranthaceae). Along with their putative progenitor species (*A. hybridus* L., *A. quitensis* H.B.K., and *A. powellii* S. Wats.) the grain amaranths are classified as part of the *A. hybridus* complex and are considered paleo-allotetraploids ($2n=4x=32$; Greizerstein and Poggio, 1994; Greizerstein and Poggio, 1995; Pal and Khoshoo, 1982) that likely arose from multiple domestication events (Mallory et al., 2008; Sauer, 1950; Rastogi and Shukla, 2013). The grain amaranths exhibit C_4 photosynthesis and thrive on marginal soils that are affected by heat and drought stress, and, occasionally, by salinity (Omami and Hammes, 2006). They are uniquely suited for subsistence agriculture and therefore, by implication, have the potential for significant impact on malnutrition (Emokaro et al., 2007). Prior to the Spanish Conquest, the grain amaranths were an important staple food crop that also played an important cultural role in the pre-Columbian civilizations. Sauer (1950, 1967; 1993) reported that the Aztec emperor Montezuma II required nearly equal annual tributes of amaranth, maize and beans. After the Spanish conquest of the New World, however, the cultivation of the grain amaranths was actively suppressed by the arriving Europeans due to their deeply rooted use in indigenous religious ceremonies (Iturbide and Gispert, 1994; Sauer, 1976; Sauer, 1993). Their use continued to decline and by the end of the 19th century they were not listed on a survey of commercial grain crops in Latin America (Sauer, 1950), though its cultivation and use persisted locally in intermediate-elevation inter-Andean valleys (*A. caudatus*) and in the Mesoamerican Highlands.

In recent years, the grain amaranths have garnered increased attention due to their nutritional qualities and in particular the nutritional value of the seed protein (Bressani et al.,

1992; Tucker, 1986). Amaranth seed is higher in fiber (8%) and fat (7 to 8%) than the grain of most cereals (Breene, 1991; Pedersen et al., 1987). Crude protein content ranges from 12.5 to 22.5% on a dry matter basis, which at the upper end of this range is more than 50% higher than the Old World grain crops of wheat, rice, and maize (Gupta and Gudu, 1991). Lysine, which is often the limiting amino acid in other cereal crops, is found in relatively high levels in amaranth seed, ranging from 0.73% to 0.84% (Bressani et al., 1987, Piskarikova et al., 2005). Seed minerals, such as iron, magnesium, and calcium are particularly high and make amaranth flour an excellent candidate for the fortification of wheat flour (Alvarez-Jubete et al., 2009). The high levels of iron found in amaranth seed has been suggested as a viable option to help fight iron-deficient anemia in developing world countries (Caselato-Sousa and Amaya-Farfan, 2012). Moreover, amaranth seed protein is gluten-free and thus represents an important protein source for sufferers of celiac disease (Alvarez-Jubete et al., 2009; Kupper, 2005; Pagano, 2006). Oil content in amaranth seeds averages about 5%, with relatively high concentrations of squalene (Belitz and Grosch, 1999). Epidemiological studies suggest that squalene reduces the risk of cancer (anti-tumor activity; Smith, 2000) and is effective in lowering cholesterol levels in humans (Berger et al., 2003; Martirosyan et al., 2007).

Many genetic resources have been developed in recent years to study the grain amaranths including genetic markers (RAPDs, Chan and Sun, 1997; SSRs, Mallory et al., 2008; SNPs, Maughan et al., 2009), genetic linkage maps (Maughan et al, 2011) and a 10x BAC library (Maughan et al., 2008). In 2014, Sunil et al. reported the draft genome sequence of an *A. hypochondriacus* land race ('Rajgira') collected locally in Karnataka, India. Unfortunately, even with 106X coverage of the genome (using short read sequences from paired-end and mate pair libraries), the Sunil et al. draft assembly for the species is highly fragmented, consisting of

491,569 contigs with an N_{50} of 1,884 bases. Efforts to scaffold the assembly produced 367,441 scaffolds, an N_{50} of 35,089 bases with 58% of the total bases being N bases (non-A, -T, -G or -C bases). Moreover, the total length of the scaffolded assembly is nearly 140% greater than the expected size of the genome, 466 Mb/C (Bennett and Smith, 1991), suggesting considerable problems with the assembly and the need for additional work to produce a high-quality reference genome.

Here we report a high-quality whole genome assembly of an agronomically important amaranth cultivar ('Plainsman', *A. hypochondriacus*) using the ALLPATHS-LG genome assembler (Gnerre et al., 2010). We further report the refinement of the genome assembly with a physical map based on BioNano™ Genomics optical mapping technology and the annotation of the genome assembly using a deeply sequenced transcriptome developed from multiple tissue types and abiotic stresses from the same cultivar used for the genome assembly. We also report the re-sequencing and comparative analysis of four additional *A. hypochondriacus* accessions, as well as one accession of *A. caudatus, A. cruentus* and *A. hybridus*. Using the assembly we identify and quantify repetitive element content in the genome and analyze the syntenic relationship between amaranth and *Beta vulgaris* L. (sugar beet), the only currently available high-quality genome assembly in the Amaranthaceae family (Dohm et al., 2014). Furthermore, using K_s (synonymous substitution rates) analysis we estimate the age of amaranth's most recent polyploidization event.

MATERIALS AND METHODS

Plant Material and Nucleic Acid Extraction. For whole genome sequencing, seeds of the 'Plainsman' cultivar (PI 558499; *A. hypochondriacus*) were kindly provided by D. Baltensperger

(Texas A&M University, USA). Seven additional accessions for re-sequencing analysis, representing all three grain amaranth species and their putative progenitor species (*A. hybridus*), were provided by David Brenner (USDA, Iowa State University, Ames, IA; Table 1). All plants for whole genome sequencing (WGS) and re-sequencing (RS) were grown in the Life Science greenhouses at Brigham Young University (Provo, UT) in 12 cm pots with one plant per pot using Sunshine Mix II (Sun Grow, Bellevue, WA) supplemented with Osmocote fertilizer (Scotts, Marysville, OH). Plants were maintained at 25°C under broad-spectrum halogen lamps with a 12-h photoperiod. For transcriptome sequencing (TS), plants were grown in a growth chamber maintained at 26°C with a 16-h photoperiod at the University of British Columbia (Okanagan campus, Kelowna, BC, Canada).

Total genomic DNA used for WGS and RS was extracted from 30 mg of freeze-dried leaf tissue from 28-day-old plants, according to the protocol devised by Sambrook et al., (1989) with modifications described by Todd and Vodkin (1996). DNA was quantified on a Nanodrop 1000 Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE). RNA was isolated for TS from a total of eight samples from the *A. hypochondriacus* cultivar 'Plainsman' representing several different tissues types, developmental stages, and abiotic stresses (Table 2). Tissue was harvested 22 days after germination from three tissue types: leaf, stem and root. Upon flowering, pooled sepals from male and female flowers were taken from the apical region of the inflorescence. Immature seeds containing globular, heart, and torpedo stage embryos were harvested and pooled as a single sample. Tissue was also harvested from mature seeds containing linear cotyledon stage embryos and from green cotyledons dissected from late stage embryos. The final sample consisted of a combination of leaf, stem, and root tissue from a subset of plants after water was withheld for 10 days and plants demonstrated loss of turgor pressure. All tissues

were collected and immediately placed in liquid nitrogen and RNA was extracted using an RNeasy Kit (Qiagen, Valencia, CA) following the manufacturer's protocol. The quality of the extracted RNA was confirmed using a BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA).

Whole Genome Library Preparation, Sequencing, and Assembly. DNA extracted from the *A. hypochondriacus* cultivar 'Plainsman' was sent to the Beijing Genomic Institute (BGI; Hong Kong, China) where one paired-end (PE) library with an insert size of 180bp, and two mate-pair (MP) libraries with insert sizes of 3kb and 6kb, were prepared and sequenced on the Illumina HiSeq platform to obtain 2x100bp reads for each library (Table 3). The generated reads were trimmed using the sliding window, quality-based trimming tool Sickle (https://github.com/najoshi/sickle), with a quality phred score cutoff of 20. The trimmed reads were then assembled using the ALLPATHS-LG assembler (Broad Institute, Cambridge, MA) using the recommended default parameters. GapCloser V.1.12, a subtool for SOAPdenovo (Short Oligonucleotide Analysis Package), was used to resolve N spacers and gap lengths produced by the ALLPATHS-LG assembler. The final assembly is hereafter referred to as the GapCloser V1.0 assembly.

Repeat Analysis. RepeatModeler v1.0.8 and RepeatMasker v.4.0.5 were used to identify and characterize repeats in the GapCloser V1.0 assembly relative to RepBase libraries (20140131; www.girinst.org). The programs use two *de novo* repeat finding programs, RECON v1.08 and RepeatScout v1.0.5, to identify repetitive elements. "One code to find them all", a Perl tool (Bailly-Bechet et al., 2014) was then used to more accurately determine the number, position, length and divergence of transposable elements in the GapCloser V1.0 assembly.

Transcriptome Library Preparation, Sequencing, and Assembly. Extracted RNA (~40 µg) was sent to BGI for library preparation and sequencing. cDNA libraries were created and then sequenced from the eight tissue types (Table 2) using 90bp PE HiSeq Illumina sequencing of 200bp inserts. ABySS V1.5.2 (Simpson et al., 2009), was used to assemble the *de novo* transcriptome. ABySS uses a k-mer sweep (k=45, 50, 55, 60, 65, 70, 80) followed by pooling of all k-mer assemblies. These pooled assemblies were then deredundified and merged with the overlap-layout-consensus assembler CAP3 (Huang and Madan, 1999). The Burrows-Wheeler Aligner (Li and Durbin, 2009) was used to align the reads back to the *de novo* assembled transcriptome to verify the validity of the assembly. The assembled transcriptome was functionally annotated using Blast2GO (B2G; Götz et al., 2011) and the NCBI non-redundant protein sequence database (28.05.2015) as a reference to provide a functional annotation for each scaffold using default parameters (*E-value* hit filter $1.0E^{-3}$, annotation cutoff 55, GO weight 5, HSP-hit coverage cut-off 20). Associated biological pathways were determined via the KEGG pathway (Ogata et al. 1999) mapping functionality using the *Viridiplantae* (nr subset; taxa:33090) database offered in the B2G program.

Gene Space and Genome Annotation. Gene space was assessed using the Core Eukaryotic Genes Mapping Approach (CEGMA; Parra et al., 2009) and *de novo* transcriptome assembly to the GapCloser V1.0 assembly using GMAP v.20141016 (Wu and Watanabe, 2005) with default parameters.

The MAKER pipeline (Cantarel et al., 2008) was used to annotate the GapCloser V1.0 assembly after masking of repetitive elements using the .consensi file generated by RepeatMasker. Evidence files, including 27,421 *Beta vulgaris* predicted gene models and their translated protein sequences from the RefBeet-1.1 assembly

(http://bvseq.molgen.mpg.de/Genome/Download/RefBeet-1.1/), the uniprot_sprot database (ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/; downloaded 3/1/15) and the *de novo* amaranth transcriptome presented here were provided as evidence for the MAKER pipeline. SNAP and Augustus were given *Arabidopsis thaliana* and *Solanum lycopersicum* as gene prediction species models, respectively.

Whole Genome Duplication Analysis. The DupPipe pipeline (Barker et al., 2008) was used to calculate the age of the most recent whole genome duplication (WGD) event in *A. hypochondriacus*. The *de novo* transcriptome assembly was clustered into gene families that showed at least 40% sequence similarity over a minimum length of 300 bp using discontiguous MegaBlast (Zhang et al., 2000). Reading frames for each sequence pair were identified by comparing all available protein sequences of plants in GenBank using BLASTx (Altschul et al., 1997). Each gene with at least 30% sequence similarity over at least 150 sites was paired with best-hit proteins. Genes that did not meet the best-hit protein criteria were removed from further analysis. Using Genewise v2.2.2 (Birney et al., 2004), each gene was aligned against its best-hit protein to determine reading frame and create a predicted amino acid sequence. MUSCLE v3.6 (Edgar, 2004) was used to align the predicted amino acid sequence for each gene pair. DNA sequences and predicted amino acid sequences were then aligned using RevTrans v1.4 (Wernersson and Pedersen, 2003). Substitutions per synonymous site (K_s) values were calculated using the maximum-likelihood method implemented in the codeml of the PAML package (Yang, 1997) under the F3-4 model (Goldman and Yang, 1994) for each duplicate gene pair. All K_s values from a single member of a duplicate pair with a K_s=0 were removed from the dataset in order to reduce the possibility that identical genes were represented in the data. Simple hierarchical clustering was used to create phylogenies for each gene family, identified as single-

linked clusters, and the nodal K_s values were determined to mitigate the multiplicative effects of multicopy gene families on K_s values.

Synteny Analysis. The syntenic relationship between *A. hypochondriacus, B. vulgaris* (RefBeet-1.1), and *Arabidopsis thaliana* (genome assembly TAIR10) was determined for scaffolds >100 kb in length using default setting of SyMap 4.2 as described by Soderlund et al. (2006).

Genetic Diversity Analysis. DNA extracted from the seven accessions selected for re-sequencing was sent to BGI for library preparation and sequencing. For each accession, one Illumina PE library with an insert size of 800bp was generated and sequenced on the HiSeq platform. The generated reads for each accession were trimmed as previously described and mapped to the GapCloser V1.0 assembly using the default parameters of the Genomic Short-read Nucleotide Alignment Program (GSNAP; Wu and Nacu, 2010). SAM files were then processed using SAMtools (Li et al., 2009) to create sorted BAM files. InterSnp, a program in the BamBam suite of programs (Page et al., 2014), was then used to call single nucleotide polymorphisms (SNPs) between re-sequenced accessions and the reference assembly, with a minimum of 10X coverage at each SNP and a minimum allele frequency of 30%. Phylogenetic analysis of the SNP genotypic data was performed using the distance (neighbor-joining) method used in the program Neighbor, also found in the BamBam suite of programs. The neighbor-joined unrooted tree was visualized using Geneious v. 8.0.5 (http://www.geneious.com). The robustness of the phylogeny was supported by bootstrap analysis using 1,000 replicates with 25% resampling.

Optical Mapping. An optical map was created using the BioNano Genomics (BNG; San Diego, CA) platform at the Bioinformatics Center at Kansas State University (Manhattan, Kansas). High molecular weight (HMW) DNA was prepared from the *A. hypochondriacus* cultivar 'Plainsman'

according to standard BNG protocols for fresh leaf tissue. The HMW DNA was then labeled according to commercial protocols using the IrysPrep Reagent Kit (BioNano Genomics, Inc). Specifically, HMW DNA was double digested by the single stranded nicking endonucleases Nb.*Bbv*CI and Nt.*Bsp*QI and labeled with a fluorescent-dUTP nucleotide analog using Taq polymerase. Nicks were ligated with Taq DNA ligase and the backbone of the labeled DNA was stained using the intercalating dye YOYO-1. The labeled DNA was then loaded into an Irys chip and linearized DNA molecules (referred to as single molecule maps) were imaged automatically using the BNG Irys system. Single molecule maps were assembled *de novo* into consensus physical maps using the BNG IrysView software package. IrysView uses an overlap-layout-consensus model with maximum likelihood to produce the longest consensus path of single molecule maps (Cao et al. 2014). The final *de novo* physical map assembly used only single molecule maps with a minimum length of 150 kb and nine labels per molecule. The p-values for the initial assembly, extension of the assembly, and chimera detection were set to $1e^{-8}$, $1e^{-9}$, and $1e^{-15}$, respectively. Hybrid scaffolds were identified using the hybrid scaffold sub-program of IrysView using a p-value of $1e^{-8}$ for the initial and final alignment and $1e^{-13}$ for the chimera detection and merging threshold.

RESULTS AND DISCUSSION

Whole Genome Sequencing and Assembly. We selected the 'Plainsman' cultivar of *A. hypochondriacus* for whole-genome shotgun sequencing (WGS) and assembly as it is the most commonly cultivated grain amaranth variety in the United States, due in part to its early maturity (~110 day), short stature and light tan seeds (Baltensperger, 1992; Brenner, 1992). Moreover, several molecular resources, including a genetic linkage map (Maughan et al, 2011) and a 10x

BAC library (Maughan et al., 2008) have previously been developed using this variety, which aided in the verification of the quality of the genome assembly reported herein.

The ALLPATHS-LG assembler utilizes a novel set of algorithms to effectively deal with repetitive sequences, low coverage regions and error correction using a k-mer sweep approach (Gnerre et al. 2010). The suggested implementation of the ALLPATHS-LG assembler requires three input data sources, including i) a 180 bp PE library at 45x coverage, ii) a short jump 3 kb MP library at 45x coverage, and iii) a long jump 6 kb MP library at 5x sequence coverage - although higher coverage confers an advantage in assembly quality. For the amaranth assembly we generated 383,396,522 (82.3x coverage) PE reads, 347,032,278 (74.5x coverage) short jump reads and 336,461,980 (72.2x coverage) long jump reads for a total of 106 Gb of sequence data (229x coverage; Table 3). The mean length of the original fragments was 97 bp with a mean G+C content of 34.7%, which is similar to most dicots (e.g., *B. vulgaris*, 34.04% G+C) and nearly identical to the 35% G+C content previously reported for BAC end sequencing in amaranth (Maughan et al. 2008).

A k-mer analysis (at k=25 scale) performed by the ALLPATHS-LG assembler predicted the genome of *A. hypochondriacus* to be 431.8 Mb which is slightly smaller than the previously flow-cytometry reported genome size of 466 Mb (Bennett and Smith, 1991). The decrease in predicted size is likely a reflection of a significant repeat fraction in the genome that is difficult to account for using a k-mer analysis. Inspection of the cumulative k-mers plot, suggests that approximately 36% of the genome is repetitive (Fig. 1A), significantly less than that predicted by RepeatModeler (see below). The k-mer spectrum plot shows no sequencing bias (often due to high G+C content) and, as expected for a mainly inbreeding species, the k-mer spectrum plot gives no indication of large-scale heterozygosity within the genome (Fig. 1B).

Prior to scaffolding, the ALLPATHS-LG assembled the short reads into 23,420 contigs with an N_{50} of 32,798 bp that totaled 357 Mb (76.6% of the predicted 466 Mb genome size). The longest contig spanned 412 kb (Table 4). Assembled contigs were then further assembled into scaffolds using the MP reads. Ninety-eight percent of the short-jump read pairs and 96.0% of the long-jump read pairs mapped to their respective length intervals of 2-4 kb and 4-8 kb producing a scaffolded genome that spanned 376.7 Mb (80.8% of the predicted genome size) in a total of 3,518 scaffolds. The N_{50} of the scaffolded genome was 371,465 bp with gaps spanning 5.1% of the sequence (N-spacers). The longest scaffold spanned 2.5 Mb. In order to better resolve the gap length between contigs, GapCloser was run on the ALLPATHS-LG assembly. GapCloser reduced the total gap length from 19.31Mb to 11.97Mb and the number of contigs in the assembly to 17,366. As expected, the number of scaffolds remained the same and the total length of the assembly was only slightly reduced (376.4 Mb, Table 4). The distribution of contigs/scaffolds lengths is shown in Figure 2a. Gap length ranged from 1 bp to 7,762 bp with a mean gap length of 865 (Fig. 2B). The amended assembly generated by GapCloser was used in all subsequent analyses (GapCloser V1.0 assembly).

Assembly Verification. A 10x BAC library (Maughan et al. 2008) was used to verify the quality of the GapCloser V1.0 assembly by i) comparing the spanning distance of the Sanger derived BAC end sequences (BES) to the GapCloser V1.0 assembly and ii) aligning all scaffolds >50 kb derived from a focused genome sequencing (FGS) assembly of 2,304 BAC clones to the GapCloser V1.0 assembly to examine coverage and base identity.

BAC end sequencing of 384 random BAC clones produced 728 reads, totaling 563 kb, of which 670 (92%) had a single significant hit (e-value $<1E^{-100}$) when aligned to the GapCloser V1.0 assembly. Of these, 316 were from paired sequences of the same BAC clone and aligned to

the same GapCloser scaffold. The average distance spanned by paired ends of BAC clones ranged from 8.4 to 206 kb with an average of 134 kb, which is very similar to the average insert size (139 kb) and range (12 to 354 kb) reported for the BAC library (Fig. 2C). To check for coverage and base identity, 2,304 random BAC clones were pooled, sequenced and assembled by Amplicon Express Inc. (http://www.ampliconexpress.com/) using FGS, a proprietary process that produces next-generation sequencing libraries of pooled BAC clones that can be assembled into high quality consensus sequences using Illumina short read sequencing data. The FGS BAC sequence assembly consisted of 15,921 scaffolds, totaling 134 Mb (28.7% of the predicted genome size) with an N_{50} of 27 kb and a maximum scaffold length of 292 kb. From the FGS assembly, we aligned all 434 scaffolds > 50 kb to the GapCloser V1.0 assembly. Of the 434 FGS-derived scaffolds, 433 (>99%) aligned with an average identity and coverage of 99.5% and 93.0%, respectively. The average coverage, while very high, is likely an underestimate since several FGS-derived scaffolds actually aligned across multiple GapCloser scaffolds and we report only the top alignment for each FGS-derived scaffold. For example, the top alignment for FGS-derived scaffold_75 (143.5 kb) is to a 97 kb end segment of GapCloser scaffold_00050 (>99% identity), but the FGS-scaffold also aligns significantly (>99% identity) to a 42 kb end segment of GapCloser scaffold_00869 (Fig. 3) – also suggesting that these GapCloser scaffolds are bridged by the FGS scaffold and should be joined. The correlated alignment of the BAC end sequence pairs and FGS scaffolds with the GapCloser V1.0 assembly supports the quality of GapCloser V1.0 assembly and suggests the potential for using the BAC library to further consolidate the assembly. A seven plate super-pooled version of the BAC library is available and was previously shown to be amendable to high-throughput BAC clone identification using the Fluidigm™ EP1 system (Maughan et al., 2012).

The completeness of the GapCloser V1.0 assembly was also assessed using a SNP linkage map developed from an interspecific (*A. caudatus* x *A. hypochondriacus*) mapping population (Maughan et al. 2011). The linkage map consisted of 411 SNPs that mapped to 16 linkage groups, believed to correspond to the haploid chromosomes found in amaranth (2n=32). BLASTn analysis, using 200 bp of flanking sequence information for each SNP, aligned 410 (99%) of the SNPs to the GapCloser V1.0 assembly with average identity and coverage of 99.2% and 99.4%, respectively, suggesting uniform coverage of the genome was achieved by the assembly. Moreover, 31 pairs of SNPs aligned to the same GapCloser scaffold which allowed us to estimate that the mean distance spanned by a centiMorgan (cM) was 152 kb (SD = 114 kb; Fig. 2D). This estimate is similar to those reported for several other plant species (e.g., *Arabidopsis thaliana*: 150 kb/cM, Heslop-Harrison, 1991) but should be used only as a reference, since genomes have "hot" and "cold" spots of recombination, often related to chromosome structure, and as such kb/cM estimates are expected to vary dramatically along the length of the chromosome due to uneven recombination rates (Fig. 2D).

Analysis of Repetitive Elements. RepeatModeler and RepeatMasker were used to identify and annotate the repetitive fraction of the GapCloser V1.0 assembly. In total, 47.7% (179 Mb) of the genome was classified as repetitive elements with an additional 1.8% (5.8 Mb) being classified as simple repeat or satellite type repeats (Table 6), with the most common di-, tri- and tetra-nucleotide simple sequence repeat (SSR or microsatellite) motif identified being $(AT)_n$, $(AAT)_n$, and $(TTTA)_n$, respectively. Previous reports on the development of SSRs for amaranth similarly report that the most common motifs are all AT-rich (Mallory et al. 2008). To date, only 179 SSRs have been characterized in amaranth, here we report the identification of 48,708 SSRs that could be targeted for development of additional SSR marker loci. Of the remaining repeat

fraction, RepeatMasker classified 15.13% (58 Mb) and 7.93% (29 Mb) of the assembly as retrotransposons and DNA transposons, leaving nearly 26% (92 Mb) classified as "unknown". This large unknown classification suggests that amaranth possesses unique repeat elements not represented in RepBase. Of the classified elements, the majority of the retrotransposons were classified as LTRs (10.94%), which can be sub-classified into copia-like (6.80%) and gypsy-like elements (3.85%) both of which are commonly found in angiosperms (Suoniemi et al., 1998). The largest subclass of DNA transposon was TcMar-Stowaway-like elements that made up nearly 2.2% of the amaranth assembly. To determine if any of the copia-, gypsy- or TcMar-Stowaway-like elements were potentially active, we plotted the percent divergence of each repeat element copy against the length ratio compared to the reference element. Copies with a divergence close to 0 and length ratio close to 1 are potentially full-length active elements; whereas those copies with increased divergence and decreased size ratios are likely degraded and inactive. In all cases, we identified copies that were putatively still active. For example, of the 39,246 TcMar-Stowaway subfamily elements examined, 1,936 (5.2%) were nearly full length (>90%) with less than 20% divergence from the reference sequence, suggesting that TcMar-Stowaway copies are likely active in the amaranth genome (Fig. 4). Mutable phenotypes, attributed to active transposition, have long been observed by amaranth breeders (personal communication, Luz Gomez-Pando, Universidad Nacional Agraria La Molina, Lima, Peru; Fig. 5) and recent work by Gaines et al. (2013) implicated transposable elements in the amplification of the 5-enolpyruvylshikimate-3-phosphate synthase gene conferring resistance to the herbicide glyphosate in *A. palmeri* (a sister taxon). Repeat sequence content within published plant genomes range dramatically, from 3% for the minute 82 Mb genome of *Utricularia gibba* (Ibarra-Laclette et al. 2013) to 85% for *Zea mays* (Schnable et al. 2009), is highly correlated with

genome size variation and is currently believed to be an important driver of genome organization and evolution (Michael, 2014). *B. vulgaris,* the only other member of the Amaranthaceae family with a published genome (Dohm et al., 2014) has a repeat content of 63%, similar to the repeat content we report for amaranth.

Gene Space. To assess the completeness of the gene space in the GapCloser V1.0 assembly we investigated the coverage of the CEGMA set of 248 core eukaryotic genes (Parra et al. 2009) in the assembly. These core eukaryotic genes (CEGs) represent a core set of highly conserved proteins found in a wide range of taxa. Of the 248 CEGs, 244 (98.4%) were identified in the assembly, of which 217 (87.5%) were considered complete with an alignment over more than 70% of their sequences, which is suggestive of a complete genome assembly (Table 7). CEGMA analysis also identified an average of 2.3 orthologs per CEG in the genome, which likely reflects a WGD event in the evolution of the amaranths (see Genome Duplication section).

Transcriptome Assembly and Functional Annotation. To facilitate annotation of the GapCloser V1.0 assembly, we created a *de novo* transcriptome assembly for the cultivar 'Plainsman' from eight different tissue and abiotic stress libraries (Table 2). Sequencing of the RNA-seq libraries generated 352,557,994 reads totaling 31.7 Gb with an average of 3.97 Mb per tissue type. ABySS, a *de novo*, parallel, paired-end sequence assembler that was designed for short reads, was used to assemble the reads into 66,370 contigs (57.3 Mb; N_{50}=1,491 bp) which were then further assembled into 65,947 scaffolds (57.3 Mb; N_{50}=1,500 bp; Table 5). The quality of the transcriptome was verified by mapping the RNA-seq reads back to the *de novo* transcriptome. A total of 97% of the reads mapped back to the assembly with 92% of the reads being properly paired, suggesting a well-assembled transcriptome. Using GMAP we investigated the relationship of the transcriptome assembly with the GapCloser V1.0 genome. Nearly 99% of

the transcriptome scaffolds aligned to the GapCloser V1.0 genome assembly with 93% having at least 80% genome coverage and 90% identity (Table 10). Of the 65,947 transcriptome scaffolds, 99% were assigned InterProScan motifs and 46% had significant BLASTx matches to known proteins in the NCBI nr protein database, with the top species alignment to *B. vulgaris* (Fig. 6A). Seven percent of the sequences with significant BLAST alignments could not be linked to Gene Ontology (GO) entries, and another 7% of the sequences with GO mapping did not surpass the quality threshold for an annotation assignment. In total, B2G annotated 21,083 (32%) sequences, and assigned enzyme codes to more than 10% of the sequences. The most dominant term seen in the biological process, molecular function and cellular component categories were oxidation-reduction process, protein binding and membrane component, respectively (Fig. 6B).

The *Amaranthus* genus contains many weedy species, collectively referred to as "pigweeds", that are among the most noxious and widespread weeds in the world. Epitomizing their weed status, the pigweeds are notorious for their ability to develop resistance to herbicides. To show the utility of the transcriptome assembly, we examined three KEGG (Kyoto Encyclopedia of Genes and Genomes) biosynthetic pathways specifically targeted by major herbicides for which resistant pigweed biotypes have been reported (Heap, 2004), including i) the biosynthesis of aromatic amino acids (phenylalanine, tryptophan, and tyrosine) which is targeted by herbicides inhibiting 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS; EC 2.5.1.19) ii) the biosynthesis of branched hydrophobic amino acids (valine, leucine, and isoleucine) which is targeted by herbicides inhibiting acetolactate synthase (ALS; EC 2.2.1.6) and iii) porphyrin and chlorophyll metabolism which is targeted by HRAC group E herbicides which inhibit protoporphyrinogen oxidase (PPO; EC 1.3.3.4). Twenty-two enzymes are required for the synthesis of the aromatic amino acids, of which 21 are found in the transcriptome

assembly, with the one missing enzyme being shikimate kinase (EC 2.7.1.71) which catalyzes the ATP-dependent phosphorylation of shikimate to form shikimate 3-phosphate and is the fifth step of the shikimate pathway. All eight enzymes necessary for the synthesis of branched amino acids were found in the transcriptome assembly as well as all sixteen enzymes necessary for porphyrin and chlorophyll metabolism (Fig. 7A-C).

Beyond the completeness of the transcriptome, the accuracy of the transcriptome can similarly be verified using herbicide gene targets. The ALS and protoporphyrinogen oxidase (PPX2L) genes were previously cloned and sequenced from 'Plainsman' (GenBank #EU024568.1 and EU024569.1; Maughan et al. 2008). The ALS gene is intronless and spans 2,010 bp. A BLASTn search of the coding sequence identified a transcript in the assembly with 99.8% identity and 100% coverage. The PPX2L gene is necessary for the production of both chlorophyll (needed for photosynthesis) and heme (needed for electron transfer chains). The gene is 11,716 bp, including 18 exons comprising 1,608 bp of coding sequence. A BLASTn search of the coding sequence identified a single transcript in the assembly with 99.9% identity and 100% coverage. While EPSPS has not been previously cloned and sequenced from *A. hypochondriacus*, it has been sequenced from a cDNA library of a resistant biotype of *Amaranthus palmeri* (FJ861243.1; Gaines et al., 2013). A BLASTn search of the coding sequence (1,556 bp) identified a single transcript in the assembly with 97.9% identity and 100% coverage. These analyses not only demonstrate accuracy of the transcriptome, but also the value of the transcriptome for gene discovery in *A. hypochondriacus* and across the Amaranthaceae family.

Gene Prediction. Annotation of the GapCloser V1.0 assembly was accomplished using the MAKER genome annotation pipeline. MAKER aligns ESTs and proteins to a genome and

creates *ab-initio* gene predictions with SNAP (Korf, 2004) and Augustus (Stanke and Waack, 2003) using evidence-based quality values. Evidence used to create the gene models consisted of our *de novo* transcriptome, the translated RefBeet-1.1 gene index from *B. vulgaris* and the uniprot_sprot database. For repeat masking, we used the amaranth specific repeats identified by RepeatMasker as described above. The MAKER pipeline identified a set of 23,059 gene predictions with 14,106 (62%) being over 1 kb in length. The mean and median transcript lengths were 1,452 bp and 1,252 bp, respectively. To assess the quality of the annotation, Annotation Edit Distance (AED) was calculated (Eilbeck et al., 2009). AED is calculated by integrating three traditional measures of annotation quality: sensitivity, specificity, and accuracy. Greater than 80% of the gene predictions in the GapCloser V1.0 assembly had AED values <0.25, which is similar to the MAKER *de novo* annotation of a benchmark region on maize chromosome 4 (Holt and Yandell, 2011) and indicative of a well-annotated genome (Fig. 8).

Genome Duplication and Synteny Mapping. *Amaranthus*, like many plant genera, has experienced a WGD event in its evolutionary history and is considered a paleo-allotetraploid, with the majority of reported species having n=16 (*A. hypochondriacus*, *A. cruentus*, *A. caudatus*, *A. quitensis*, *A. edulis* L., *A. powellii* and *A. retroflexus* L.) or n=17 (*A. tricolor* L. and *A. spinosus* L.; http://www.data.kew.org). The DupPipe pipeline was used to determine when the most recent polyploidization event occurred using synonymous nucleotide substitutions per synonymous site (K_s) divergence between 860 duplicate gene pairs. The K_s distribution clearly identified a single duplication event that peaked at $K_s = 0.55$ (Fig. 9). Based on this modal K_s value, we estimate the timing of WGD event to have occurred in amaranth between 36.7 – 67.9 MYA, depending on whether an arabidopsis-based synonymous mutation rate of 1.5×10^{8} (Koch et al., 2000) or a core eukaryotic-based rate 8.1×10^{9} (Lynch and Conery, 2000) is assumed. *B.*

vulgaris, a member of the Betoideae subfamily of Amaranthaceae is a diploid species (n=9) that is believed to have diverged from amaranth approximately 35.4 – 48.6 MYA (Hohmann et al., 2006). Accordingly, we infer that the WGD event in amaranth occurred after the Betaoideae subfamily diverged, between 36.7 – 48.6 MYA, and was likely followed by chromosomal loss.

The shared ancestry between amaranth and *B. vulgaris* was also seen in the significant level of synteny observed between the two genomes using SyMap 4.2 (Soderlund et al., 2006). Aligning all amaranth scaffolds with lengths over 100 kb (943 scaffolds totaling 324 Mb) to the nine chromosomes of *B. vulgaris* identified a total of 35,948 hits in 778 syntenic and collinear blocks represented on 554 scaffolds (208 Mb). The 554 scaffolds covered 91% of *B. vulgaris* chromosomes (Fig. 10) demonstrating a high level of synteny between the two genomes. For comparison, only 40% and 22% of the arabidopsis and rice genomes aligned with the amaranth scaffolds. A total of 34 (6.1%) scaffolds showed synteny split across two *B. vulgaris* chromosomes while the remaining 520 mapped to unique loci on the *B. vulgaris* chromosomes. Interestingly, 57% of the *B. vulgaris* genome was double covered by amaranth scaffolds, where two or more scaffolds map to the same *B. vulgaris* chromosomal region. Considering amaranth's paleotetraploid evolution, it is likely that these scaffolds represent homoelogous regions of the amaranth genome.

Genetic Diversity Analysis. As a preliminary effort to characterize the genetic diversity within and among the amaranth grain species and their putative ancestor (*A. hybridus*) we re-sequenced four accessions of *A. hypochondriacus* and one accession each of *A. caudatus, A. cruentus,* and *A. hybridus* (Fig. 11). The re-sequencing yielded between 15.6 and 15.9 Gb of sequence data per accession (Table 8). The reads generated were then mapped back to the GapCloser V1.0 assembly and analyzed using InterSnp (Page et al. 2014) to identify SNPs in the seven

accessions. A total of 7,495,570 putative SNPs were identified across all seven accessions using a minimum of 10X coverage and a minimum SNP frequency in the panel of 30%. Percent heterozygosity was also calculated for each of the seven accessions and as was anticipated, the weedy species, *A. hybridus,* was the most heterozygous (8.6%), with the cultivated species ranging from 3.7 – 7.6%, which corresponds well with the low outcrossing rate reported for the grain amaranths (10.4 – 10.9%; Agong and Ayiecho, 1991).

We divided the seven accessions into two groups to investigate i) intraspecific genetic diversity within *A. hypochondriacus* and ii) interspecific diversity among the grain amaranth species and *A. hybridus*. In the intraspecific group, 1,760,433 putative SNPs were identified, with relatively few autapomorphic SNPs identified in the accessions collected from India (PI481125; 37,123 SNPs), Nepal (PI619259; 43,542 SNPs), and Pakistan (PI540446; 133,221 SNPs) in comparison to the 685,639 autapomorphic SNPs identified in the accession collected in Mexico (PI511731; Fig. 12A). These results were not unexpected as PI481125, PI619259, and PI540446 were previously shown to be much more closely related to each other than to PI511731 (Maughan et al., 2011), and is likely a reflection of a genetic bottleneck that occurred at the time of dispersal of amaranths during colonial times to South Asia from Mexico. Mexico is believed to be the center of diversity and domestication for *A. hypochondriacus* (Sauer, 1967). As expected, the interspecific group comparisons identified significantly more putative SNPs (7,184,636; Fig 12B). For comparison, the only other large-scale SNP discovery effort reported in the literature for grain amaranths used genomic reduction and 454-pyrosequencing to identify a total of 27,658 SNPs (Maughan et al. 2009).

There are several hypotheses that have been proposed for the evolutionary origins of the grain amaranths. The first hypothesis is based on geographic separation and suggests that the

grain species arose independently, specifically *A. caudatus* from *A. quitensis* in South America, *A. cruentus* from *A. hybridus* in Central America and *A. hypochondriacus* from *A. powellii* in Mexico (Sauer, 1967, 1976). The second hypothesis, based on plant morphology, suggests that *A. cruentus* arose from *A. hybridus*, which in turn hybridized with *A. powellii* and *A. quitensis* to give rise to *A. hypochondriacus* and *A. caudatus*, respectively. A third hypothesis, based on recent studies using molecular markers showed that *A. hybridus* is polyphyletic while the grain species are monophyletic, suggesting that all three grain amaranth species arose directly from *A. hybridus* in multiple independent domestication events (Mallory et al., 2008). While the number of accessions included in our RS analysis is much too small to make definitive evolutionary conclusions, the neighbor-joining analysis of our SNP data mirrors those of Mallory et al. (2008) and Maughan et al. (2011), supporting the designation of *A. hybridus* as the progenitor species of the grain amaranths and the close relationship between *A. caudatus* and *A. hypochondriacus* (Kietlinski et al. 2014). We note that the close relationship between the *A. hybridus* (PI605351) and *A. cruentus* (PI477913) accessions, as seen in both the dendrogram (Fig. 13) and by significantly greater number of shared SNPs (1,422,152; Fig. 12B), is explained by the polyphyletic clustering of *A. hybridus* with the grain amaranths, where this specific *A. hybridus* accession is grouped with the *A. cruentus* clade as previously reported by Mallory et al. (2008) – indeed, if other *A. hybridus* accessions had been included in the analysis we would have expected *A. hybridus* accessions aligning with each of the grain amaranth clades. The tree also justifies the assertion that PI481125 was originally misclassified as *A. caudatus* in the USDA GRIN system and should be reclassified as *A. hypochondriacus,* as it clearly grouped with the other *A. hypochondriacus* accessions (and is treated as such in this research). Clearly, the GapCloser V1.0 assembly presented here represents an important genomic resource necessary

for larger phylogenetic investigations that will finely dissect the taxonomic origins of the grain amaranths.

Optical Map. The BioNano Genomics (BNG) Irys System was used to create an optical map of the *A. hypochondriacus* genome. A total of 169 Gb (363x coverage) of data were generated on the Irys Instrument. After filtering out low quality single molecule maps (read length less than 150 kb or label density less than nine per molecule), a total of 80.7 Gb (~173x coverage) of data were included in the final *de novo* assembly of the optical map. The final optical map assembly consisted of 619 genome maps that spanned 340 Mb (73.0% genome coverage). To make a hybrid assembly of the optical genome maps and the GapCloser V1.0 assembly, all GapCloser V1.0 scaffolds greater than 20 kb in length with a minimum of five labels sites (1,419 scaffolds), as determined by an *in silico* digestion, were aligned to the optical genome. Of the 619 optical genome maps, 494 (80%) aligned to the 1,419 *in silico* digested scaffolds with a unique alignment length of 221.7 Mb. The hybrid assembly collapsed 343 GapCloser V1.0 scaffolds into 241 hybrid scaffolds, reducing the final number of scaffolds in the hybrid assembly to 3,416 (3,175 scaffolds and 241 hybrid scaffolds). The hybrid assembly nearly doubled the N_{50} of the assembly (696,622 bp), with the longest scaffold spanning 5,997,829 bp and a total length in the hybrid assembly of 429 Mb (92% of the predicted genome size). As expected the %N in the hybrid assembly increased (15.08%), while the L_{50} was reduced to 147 (Table 4).

The validity of the hybrid assemblies were verified, when possible, using genetic markers and/or BAC end sequences. For example, two BioNano optical maps (34 and 1378) bridged three GapCloser V1.0 scaffolds [scaffold #s: 00027 (1.4 Mb), 00071 (0.8 Mb) and 00378 (0.3 Mb)] producing a single hybrid molecule spanning 2.72 Mb (hybrid scaffold_43; Fig. 14). The bridging of the scaffolds was supported by three SNPs (AM23401, AM20979, AM22341) that

were previously shown to be linked together on linkage group 8 (Maughan et al. 2011). The SNPs were linked by a total genetic distance of 13.0 cM, which according to our 152 kb/cM estimate (see Assembly Verification) should represent a total nucleotide distance of 1.98 Mb (AM20979 – AM22341), which is very close to the 1.84 Mb separating the SNP sequences in the hybrid scaffold. The bridging of scaffolds was further supported by paired BES of BAC clone PBa0043bB23, which uniquely (>99% identity) mapped across the gap between scaffolds 00027 and 00071 (Fig 14). Within the hybrid scaffold, the BES are separated by 216 kb which is within the size range reported for the BAC clones library. In no cases did we see disagreement with the linkage map or paired BES with the hybrid assemblies produced by IrysView.

While the IrysView hybrid assembler significantly increased the overall length and N_{50} of the assembly, it had only a minimal effect on the total number of scaffolds (reducing the count by 102 scaffolds). This is likely a reflection of the distribution of scaffolds sizes in the GapCloser V1.0 assembly (few very large scaffolds and overwhelming numbers of small scaffolds) and the inherent limitations of the BioNano technology, which require large molecules with high label densities for hybrid assembly. Indeed, nearly 60% of the GapCloser V1.0 scaffolds are too small or lack the necessary *in silico* label density necessary to be included into the hybrid assembly analysis. Thus only the largest scaffolds tend to benefit from the BioNano hybrid assembly. The utility of the technology would likely be expanded with the incorporation of additional nicking enzymes and/or dual labeling with a second fluorescent-dUTP nucleotide analog – one for each nicking enzyme. Dual labeling would enhance the restriction pattern recognition and the ability of the IrysView software to utilize smaller scaffolds in the hybrid assembly – even in a repeat rich genome.

CONCLUSIONS

The GapCloser V1.0 assembly reported here consists of 104x fewer scaffolds (3,518 vs. 367,441), has an N_{50} over 10x larger (371 kb vs. 35 kb) and is comprised of nearly 95% fewer N-spacers (3.18% vs. 57.6%) than a previously published draft of the amaranth genome (Sunil et al., 2014) and thus represents a significant step forward towards development of a full length, reference quality genome sequence for the amaranths. The greater contiguity and completeness of this assembly should facilitate higher quality and more accurate analysis of genome composition (gene number, gene function, repeat fraction), genetic diversity and the investigation of phylogeny. Indeed the value of a high quality genome assembly can be seen in the dramatic difference of the repeat faction of the genome predicted by the highly fragmented early draft assembly (13.8%) versus the GapCloser V1.0 assembly (47.7%) reported here.

We report the first optical map for the species, which increased the N_{50} of the assembly by nearly two-fold. Optical maps are not only valuable for consolidating genome assemblies, but are of primary value for identifying structural rearrangement at the chromosome level (English et al., 2015). One of the earliest events associated with speciation are postzygotic barriers, which are often chromosomal rearrangements (e.g., translocations, inversions, chromosome fusions/fission) that disrupt mitotic/meiotic division leading to hybrid breakdown and sterility (Stelly et al. 1990; Morikawa and Leggett, 1996). A classic textbook example of hybrid breakdown is the mule, where sterility is caused by meiotic dysfunction due to aberrant chromosomal pairing of differentiated chromosomes and a major chromosome fusion/fission *(E. caballus, 2n=64; E. asinus, 2n=62).* The speciation of the grain amaranths is an evolutionarily new event (Muller et al. 2005) and as such they represent a unique opportunity to characterize the initial phases of genome differentiation associated with postzygotic speciation. Indeed,

crosses between the grain amaranths and their wild relatives show varying degrees of hybrid breakdown. The development and comparison of optical maps across the genus should allow us to characterize, at a resolution never seen before, the early patterns of structural genome evolution associated with speciation.

Further improvements to the assembly could be made by the use of i) Pacific Biosciences® long read technology which would facilitate the consolidation of smaller scaffolds as well as reduce the number of gaps within the larger scaffolds; and ii) genome conformation capture (aka Hi-C) technology, such as described by Putnam et al. (2015) where read pairs are generated by proximity ligation of DNA based on *in vitro* reconstituted chromatin. These very long jumping libraries can span gaps that are not possible to be spanned using current sequencing technologies (e.g., VNTRs, centromeres, etc.) and increase the scaffold contiguity of assemblies as well as provide haplotype-phasing information. Notwithstanding, this assembly is the first high-quality draft genome sequence available for the study of amaranth and will be an invaluable resource for further research.

REFERENCES

Agong, S.G., and P.O. Ayiecho. 1991. The rate of outcrossing in grain amaranths. Plant Breeding. 107(2):156-160.

Altschul, S.F., T.L Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. A new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Alvarez-Jubete, L., M. Holse, Å. Hansen, E.K. Arendt, and E. Gallagher. 2009. Impact of baking on vitamin E content of pseudocereals amaranth, quinoa, and buckwheat. Cereal Chem. 865:511-515.

Bailly-Bechet, M., A. Haudry, and E. Lerat. 2014. "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. Mobile DNA. 5:13.

Baltensperger, D.D., L.E. Weber, and L.A. Nelson. 1992. Registration of 'Plainsman'grain amaranth. Crop Sci. 32(6):1510-1511.

Barker, M.S., N.C. Kane, M. Matvienko, A. Kozik, W. Michelmore, S.J. Knapp, and L.H. Rieseberg. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal paraellel patterns of duplicate gene retention after millions of years. Mol. Biol. Evol. 25:2445-2455.

Belitz, I.H.D., and I.W. Grosch. 1999. Cereals and Cereal Products. p. 631-692. Food Chem. Springer Berlin Heidelberg.

Bennett, M.D., and J.B. Smith. 1991. Nuclear DNA amounts in angiosperms. Philos. T. Roy. Soc. B. 334:309-345.

Berger, A., G. I. Monnard, F. Dionisi, D. Gumy, K.C. Hayes, and P. Lambelet. 2003. Cholesterol-lowering properties of amaranth flakes, crude and refined oils in hamsters. Food Chem. 81:119-124. doi:10.1016/S0308-8146(02)00387-4

Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and genomewise. Genome Res. 14(5): 988-995.

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

Brenner, D.M. 1992. The Plainsman Story. Legacy. 5(1):12-13.

Bressani, R., J.M. Gonzales, J. Zuniga, M. Breuner, and L.G. Elias. 1987. Yield, selected chemical composition and nutritive value of 14 selections of amaranth grain representing four species. J. Sci. Food Agric. 38:347-356. doi:10.1002/jsfa.2740380407

Bressani, R., A. Sánchez-Marroquín, and E. Morales. 1992. Chemical composition of grain amaranth cultivars and effects of processing on their nutritional quality. Food. Rev. Int. 8:23-49. doi:10-1080/87559129209540928

Cao, H., A.R. Hastie, D. Cao, E.T. Lam, Y. Sun, H. Huang, X. Liu, et al. 2014. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. GigaScience. 3:34.

Cantarel, B.L., I. Korf, S.M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A.S. Alvarado, M. Yandell. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188-196.

Caselato-Sousa, V.M., and J. Amaya-Farfán. 2012. State of knowledge on amaranth grain: a comprehensive review. J. Food Sci. 77:R93-R104.

Chan, K.F., and M. Sun. 1997. Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of Amaranthus. Theor. Appl. Genet. 95:865-873. doi:10.1007/s10592-008-9511-7

Dohm, J.C., A.E. Minoche, D. Holtgräwe, S. Capella-Gutiérrez, F. Zakrzewski, H. Tafer, O. Rupp, et al. 2014. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature. 505:546-549.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nuclelic Acids Res. 32:1792-1797.

Eilbeck, K., B. Moore, C. Holt, and M. Yandell. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 10:67.

Emokaro, C.O., P.A. Ekunwe, and A. Osifo. 2007. Profitability and production constraints in dry season amaranth production in Edo South, Nigeria. Int. J. Food Agric. Environ. 5:281–283.

English, A. C., W.J. Salerno, O.A. Hampton, C. Gonzaga-Jauregui, S. Ambreth, D.I. Ritter, ... and R.A. Gibbs. 2015. Assessing structural variation in a personal genome—towards a human reference diploid genome. BMC Genomics. 16(1):286.

Gaines, T.A., A.A. Wright, W.T. Molin, L. Lorentz, C.W. Riggins, P.J. Tranel, R. Beffa, P. Westra, S.B. Powles. 2013. Identification of genetic elements associated with EPSPS gene amplification. PloS One. 8:e65819.

Götz, S., R. Arnold, P. Sebastián-León, S. Martín-Rodríguez, P. Tischler, M.A. Jehl, J. Dopazo, T. Rattei, A. Conesa. 2011. B2G-FAR, a species-centered GO annotation repository. Bioinformatics. 27:919-924.

Gnerre, S., I. MacCallum, D. Przybylski, F.J. Ribeiro, J.N. Burton, B.J. Walker, T. Sharpe, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. P. Natl. Acad. Sci. 108:1513-1518.

Goldman, N., and Z. Yang. 1994. A codon based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725-736.

Greizerstein, E.J., and L. Poggio. 1994. Karyological studies in grain Amaranths. Cytologia (Tokyo). 59: 25-30.

Greizerstein, E.J., and L. Poggio. 1995. Meiotic studies of spontaneous hybrids of *Amaranthus*: Genome analysis. Plant Breed. 114:448-450. doi:10.1111/j.1439-0523.1995.tb.00830.x

Gupta, V.K., and S. Gudu. 1991. Interspecific hybrids and possible phylogenetic relations in grain amaranths. Euphytica. 52:33-38.

Heap, I.M. 2014. Herbicide resistant weeds (pp. 281-301). Springer Netherlands.

Heslop-Harrison, J. S. 1991. The molecular cytogenetics of plants. J. Cell Sci. 100(1):5-21.

Hohmann, S., J.W. Kadereit, and G. Kadereit. 2006. Understanding Mediterranean-Californian disjunctions: molecular evidence from Chenopodiaceae-Betoideae. Taxon. 67-78.

Holt, C., and M. Yandell. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics. 12(1):491.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868-877.

Ibarra-Laclette, E., E. Lyons, G. Hernández-Guzmán, C.A. Pérez-Torres, L. Carretero-Paulet, T.H. Chang, ... and L. Herrera-Estrella. 2013. Architecture and evolution of a minute plant genome. Nature. 498(7452): 94-98.

Iturbide, G.A., and M. Gispert. 1994. Grain amaranths (*Amaranthus* spp.). p. 93-101. *In* J. E. Hernandez-Bermejo and J. Leon (Ed.) Neglected Crops: 1492 from a different perspective. FAO, Rome.

Kietlinski, K.D., F. Jimenez, E.N. Jellen, P.J. Maughan, S.M. Smith, and D.B. Pratt. 2014. Relationships between the Weedy (*Amaranthaceae*) and the Grain Amaranths. Crop Sci. 54:220-228.

Koch, M.A., B. Haubold, and T. Mitchell-Olds. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (*Brassicaceae*). Mol. Biol. Evol. 17:1483-1498.

Korf, I. 2004. Gene finding in novel genomes. BMC Bioinformatics. 5:59.

Kupper, C. 2005. Dietary guidelines and implementation for celiac disease. Gastroenterology. 128:S121-S127.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25:1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. MArth, G. Abecasis, and R. Durbin. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078-2079.

Loots, G. G., and I. Ovcharenko. 2008. Mulan. Comparative Genomics. Humana Press. pp. 237-253.

Lynch, M., and J.S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151-1155.

Mallory, M.A., R.V. Hall, A.R. McNabb, D.B. Pratt, E.N. Jellen, and P.J. Maughan. 2008. Development and characterization of microsatellite markers for the grain amaranths. Crop Sci. 48:1098-1106. doi:10.2135/cropsci2007.08.0457

Martirosyan, D.M., L.A. Miroshnichenko, S.N. Kulakova, A.V. Pogojeva, and V.I. Zoloedov. 2007. Amaranth oil application for coronary heart disease and hypertension. Lipids Health Dis. 6:1. doi:10.1186/1476-511X-6-1.

Maughan, P.J., N. Sisneros, M.Z. Luo, D. Kudrna, J.S.S. Ammiraju, and R.A. Wing. 2008. Construction of an *Amaranthus hypochondriacus* bacterial artificial chromosome library and genomic sequencing of herbicide target genes. Crop Sci. 48:S85-S94.

Maughan, P.J., S.M. Yourstone, E.N. Jellen, and J.A. Udall. 2009. SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. Plant Gen. 2:260-270. doi:10.3835/plantgenome2009.08.0022

Maughan, P.J., S.M. Smith, D.J. Fairbanks, and E.N. Jellen. 2011. Development, characterization, and linkage mapping of single nucleotide polymorphisms in the grain amaranths (*Amaranthus* spp.). Plant Gen. 4:92-101.

Maughan, P.J., S.M. Smith, and J.A. Raney. 2012. Utilization of super BAC pools and fluidigm access array platform for high-throughput BAC clone identification – proof of concept. J. Biomed. Biotech. doi:10.1155/2012/405940

Michael, T. P. 2014. Plant genome size variation: bloating and purging DNA. Briefings in functional genomics. 13(4): 308-317.

Morkiawa, T., and J.M. Leggert. 1996. Cytological and morphological variations in wild populations of *Avena canariensis* from the Canary Islands. Genes Genet Syst. 71(1):15-21.

Müller, K., and T. Borsch. 2005. Phylogenetics of Amaranthaceae based on matK/trnK sequence data: evidence from parsimony, likelihood, and Bayesian analyses. Ann Mo Bot Gard. 66-102.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids Res. 27(1):29-34.

Omami, E.N., and P.S. Hammes. 2006. Interactive effects of salinity and water stress on growth, leaf water relations, and gas exchange in amaranth (*Amaranthus* spp.). New Zeal. J. Crop Hort. 34:33-44.

Pagano, A.E. 2006. Whole grains and the gluten-free diet. Pract. Gastroenterol. 30:66.

Page, J.T., Z.S. Liechty, M.D. Huynh, and J.A.Udall. 2014. BamBam: genome sequence analysis tools for biologists. BMC research notes. 7(1):829.

Pal, M., R.M. Pandey, and T.N. Khoshoo. 1982. Evolution and improvement of cultivated amaranths. J. Hered. 73:353-356.

Parra, G., K. Bradnam, Z. Ning, T. Keane, and I. Korf. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37:289-297.

Pedersen, B., L.S. Kalinowski, and B.O. Eggum. 1987. The nutritive value of amaranth grain (*Amaranthus caudatus*). Plant Food Hum. Nutr. 36:309-324.

Piskarikova, B., S. Kracmar, I. Herzig. 2005. Amino acid contents and biological value of protein in various amaranth species. Czech J. Anim Sci. 50(4):169-174.

Putnam, N. H., B. O'Connell, J.C. Stites, B.J. Rice, A. Fields, P.D. Hartley, ... and R.E. Green. 2015. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. arXiv preprint arXiv:1502.05331.

Rastogi, A., and S. Shukla. 2013. Amaranth: A new millennium crop of nutraceutical values. Crit. Rev. Food Sci. 53:109-125.

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

Sauer, J.D. 1950. The grain amaranths: A survey of their history and classification. p. 632. Ann. Mo. Bot. Gard.

Sauer, J.D. 1967. The grain amaranths and their relatives: A revised taxonomic and geographic survey. Ann. Mo. Bot. Gard. 54:103-137. doi:10.2307/2394998

Sauer, J.D. 1976. Grain amaranths. p. 4-7. *In* Evolution of Crop Plants. N.W. Simmonds (ed.). Longman Group, London, UK.

Sauer, J.D. 1993. Amaranthaceae: Amaranth family. p. 9-14. Historical Geography of Crop Plants: A Select Roster. CRC Press, Boca Raton, FL.

Schnable, P. S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, ... and M. Cordes. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326(5956): 1112-1115.

Simpson, J.T., K. Wong, S.D. Jackman, J.E. Schein, S.J. Jones, and I. Birol. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. 19:1117-1123.

Smith, T.J. 2000. Squalene: Potential chemopreventive agent. Expert Opin. Invest. Drugs. 9:1841–1848.

Soderlund, C., W. Nelson, A. Shoemaker, and A. Paterson. 2006. SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Res. 16:1159-1168.

Stanke, M., and S. Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 19:ii215-ii225.

Stelly, D. M., K.C. Kautz, and W.L. Rooney. 1990. Pollen fertility of some simple and compound translocations of cotton. Crop Sci. 30(4):952-955.

Sunil, M., A.K. Hariharan, S. Nayak, S. Gupta, S.R. Nambisan, R.P. Gupta, B. Panda, B. Choudhary, and S. Srinivasan. 2014. The draft genome and transcriptome of *Amaranthus hypochondriacus*: A C4 dicot producing high-lysine edible pseudo-cereal. DNA R. dsu021.

Suoniemi, A., J. Tanskanen, and A.H. Schulman. 1998. Gypsy-like retrotransposons are widespread in the plant kingdom. Plant. J. 13:699-705.

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

Tucker, J.B. 1986. Amaranth: The once and future crop. Bioscience. 36:9-13. doi:10.2307/1309789

Wernersson, R., and A.G. Pedersen. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31:3537-3539.

Wu, T.D., and S. Nacu. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 26:873-881.

Wu, T.D., and C.K. Watanabe. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 21:1859-1875.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555-556.

Zhang,Z., S. Schwartz, L. Wagner, and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7:203-214.

TABLES AND FIGURES

Table 1. *Amaranthus* accessions used for genome and transcriptome sequencing and re-sequencing analysis.

Name	Species	Geographical Location[1]	Analysis Type[2]
PI558499[3]	*Amaranthus hypochondriacus L.*	Nebraska, USA	WGS/TS
PI511731	*A. hypochondriacus*	Mexico	RS
PI540446	*A. hypochondriacus*	Pakistan	RS
PI619259	*A. hypochondriacus*	Nepal	RS
PI481125[4]	*A. hypochondriacus*	India	RS
PI642741[5]	*Amaranthus caudatus L.*	Bolivia	RS
PI477913	*Amaranthus cruentus L.*	Mexico	RS
PI605351	*Amaranthus hybridus L.*	Greece	RS

[1]All Origin information is derived from the Germplasm Resources Information Network (http://www.ars-grin.gov/npgs/). Several accessions were collected in the Old World although all originated in the Americas according to Sauer (1967)

[2]WGS: Whole genome sequencing; TS: Transcriptome sequencing; RS: Re-sequencing

[3]cv. 'Plainsman'

[4]Reclassified here as *A. hypochondriacus* based on Maughan et al. 2010 and data reported herein.

[5]PI 481125 and PI 642741 were the parents of the mapping population reported by Maughan et al. (2011).

Table 2. Tissues and treatment types, total number of reads and bases sequenced in the transcriptome analysis.

Tissue Details[1]	Total Number of Paired End Reads[2]	Total Number of Bases
Floral tissue (sepals from both male and female flower)	43,002,862	3,870,257,580
Leaf tissue	44,090,140	3,968,112,600
Root tissue	45,313,542	4,078,218,780
Stem tissue	44,325,240	3,989,271,600
Water stressed tissue sample (mixed- root, stem, and leaf)	43,220,952	3,889,885,680
Immature seeds	43,536,684	3,918,301,560
Mature seeds	44,485,024	4,003,652,160
Green Cotyledon (no perisperm, no seed coat)	44,583,550	4,012,519,500
Total:	352,557,994	31,730,219,460

[1]All tissues were derived from the *A. hypochondriacus* cultivar 'Plainsman'.
[2]Illumina HiSeq PE with a nominal insert size of 180 bp

Table 3. Total number of reads and bases sequenced in the whole genome sequencing analysis.

Library Details[1]	Total Number of Paired End Reads	Total Number of Bases	Coverage (X)[2]
180 bp PE	383,396,522	38,339,652,200	82.3x
3 kb MP	347,032,278	34,703,227,800	74.5x
6 kb MP	336,461,980	33,646,198,000	72.2x
Total:	1,066,890,780	106,689,078,000	228.9x

[1]PE: Paired-End; MP: Mate Pair
[2]Based on a predicted genome size of 466 Mb/C per Bennett and Smith (1991).

Table 4. Statistical summary of WGS assemblies of the amaranth variety 'Plainsman'.

	ALLPATHS-LG	GapCloser V1.0	Hybrid Assembly
Contigs	23,420	17,366	13,569
Total Size of Contigs (bp)	357,417,249	364,448,437	364,364,087
Longest Contig (bp)	301,117	412,940	581,624
N50 Contig Length (bp)	32,798	44,493	64,960
Scaffolds	3,518	3,518	3,416
Total Size of Scaffolds (bp)	376,726,029	376,423,229	429,032,159
Longest Scaffold (bp)	2,519,536	2,519,077	5,997,829
N50 Scaffold Length (bp)	371,465	370,786	696,622
L50 Scaffold Count	243	243	147
%N (%)	5.13	3.18	15.08
Total Gap Length (bp)	19,308,780	11,974,792	64,698,049
Number of Gaps	19,902	13,848	14,182
G+C Content (%)	33	33	33

Table 5. Statistical summary of de novo transcriptome assembly generated by ABySS from the eight tissue types listed in Table 2.

Contigs	66,370
Max Contig (bp)	12,410
Mean Contig (bp)	864
Contig N50 (bp)	1,491
Total Contig Length (bp)	57,327,871

Scaffolds	65,947
Max Scaffold (bp)	12,410
Mean Scaffold (bp)	870
Scaffold N50 (bp)	1,500
Total Scaffold Length (bp)	57,347,461
Assembly G+C (%)	39.05
Total Gap Length	19,590
Max Gap (bp)	522
Mean Gap (bp)	46

Table 6. Organization of repetitive elements in the 'Plainsman' genome.

Type of Element	Number of Elements	Number of bp masked	Genome (%)
DNA Transposon	**107,125**	**29,836,597**	**7.93%**
CMC-EnSpm	8,479	3,153,793	0.84%
Ginger	75	40,791	0.01%
Kolobok-Hydra	685	179,016	0.05%
MULE-MuDr	2,118	962,977	0.26%
MuLE-MuDR	5,466	2,867,477	0.76%
PIF-Harbinger	3,273	1,108,934	0.29%
TcMar-Mogwai	408	339,096	0.09%
TcMar-Stowaway	39,785	8,030,444	2.13%
TcMar-Tigger	604	208,938	0.06%
hAT-Ac	23,383	7,060,694	1.88%
hAT-Tag1	4,826	1,007,541	0.27%
hAT-Tip100	10,302	2,215,516	0.59%
Helitron	2,424	1,214,630	0.32%
Unclassified	5,297	1,446,750	0.38%
Retrotransposon	**120,990**	**56,956,858**	**15.13%**
LINE[1]	30,302	13,624,217	3.63%
CRE-II	969	784,983	0.21%
Jockey	164	59,542	0.02%
L1	9,083	7,219,515	1.92%
R1	189	236,465	0.06%
R2	489	70,241	0.02%
RTE-BovB	19,408	5,253,471	1.40%
LTR[2]	76,022	41,203,587	10.94%
Copia	48,929	25,595,527	6.80%
Gypsy	19,493	13,738,422	3.65%
Ngaro	404	275,044	0.07%
Unclassified	7,196	1,594,594	0.42%
SINE[3]	14,666	2,129,054	0.56%
Alu	233	93,585	0.02%
RTE	2,283	257,995	0.07%
tRNA	3,136	515,972	0.14%
tRNA-Core	576	129,882	0.03%
tRNA-RTE	7,529	1,064,770	0.28%
Unclassified	909	66,850	0.02%
Unknown	**427,321**	**92,503,055**	**24.58%**
Total	**655,436**	**179,296,510**	**47.65%**
Satellite	151	43,211	0.01%
Simple Repeat	99,234	5,731,032	1.52%

[1]LINE, Long Interspersed Nuclear Element
[2]LTR, Long Terminal Repeat
[3]SINE, Short Interspersed Nuclear Element

Table 7. CEG Mapping Approach using 248 CEGs to measure the completeness of the GapCloser V1.0 assembly.

	# of Proteins	% Identified[1]	# of Orthologs[2]
Complete[3]	217	88	1.77
Group 1 (n=66)[4]	56	85	1.48
Group 2 (n=56)	49	88	1.80
Group 3 (n=61)	51	84	1.86
Group 4 (n=65)	61	94	1.95
Partial	244	98	2.33
Group 1	64	97	2.03
Group 2	55	98	2.31
Group 3	61	100	2.51
Group 4	64	98	2.48

[1] % Identified equals the percentage of 248 CEGs present

[2] # of Orthologs equals the average number of orthologs per CEG

3 Complete represents the sequences that aligned over more than 70% of their sequence while partial shows all sequences that aligned regardless of coverage of the sequence.

[4]Group 1 represents the least conserved of the 248 CEGs, with the degree of conservation increasing in subsequent groups through Group 4.

Table 8. Total number of reads and bases sequenced from the seven accessions in re-sequencing.

Library Details[1]	Total Number of Paired End Reads	Total Number of Bases	Coverage (X)[2]
PI477913	158,935,806	15,893,580,600	34.1x
PI481125	158,212,794	15,821,279,400	34.0x
PI511731	156,230,518	15,623,051,800	33.5x
PI540446	156,496,790	15,649,679,000	33.6x
PI605351	156,962,464	15,696,246,400	33.7x
PI619259	157,743,254	15,774,325,400	33.9x
PI642741	157,012,184	15,701,218,400	33.7x

[1]All reads were generated from 800 bp PE libraries.

[2]Based on a predicted genome size of 466 Mb/C per Bennett and Smith (1991).

Table 9. Annotation statistics of the GapCloser V1.0 assembly produced by MAKER.

Number of Proteins (aa)	**23,059**
Total Size of Sequence	8,991,366
Longest Protein	5,322
Number of Proteins >1K	994
Mean Size	390
Median Size	320
Number of Transcripts (nt)	**23,059**
Total Size of Sequence	33,470,982
Longest Transcript	16,139
Number of Transcripts >1K	14,106 (62%)
Mean Size	1,452
Median Size	1,252
N50	1,901
L50	5,889
%A	29.51
%C	18.73
%G	21.85
%T	29.92
%N	0

Table 10. Alignment of the transcriptome assembly to the GapCloser V1.0 assembly.

			At Least One Mapping		Unique Mapping	
Total Transcripts	**Query Coverage**	**Query Identity**	**Total**	**%**	**Total**	**%**
65,947	≥ 0%	≥ 90%	65,164	98.8	51,570	79.1
	≥ 90%	≥ 90%	58,987	90.5	50,372	77.3
	≥ 80%	≥ 90%	60,829	93.3	50,601	77.7
	≥ 50%	≥ 90%	64,089	98.4	51,215	78.6

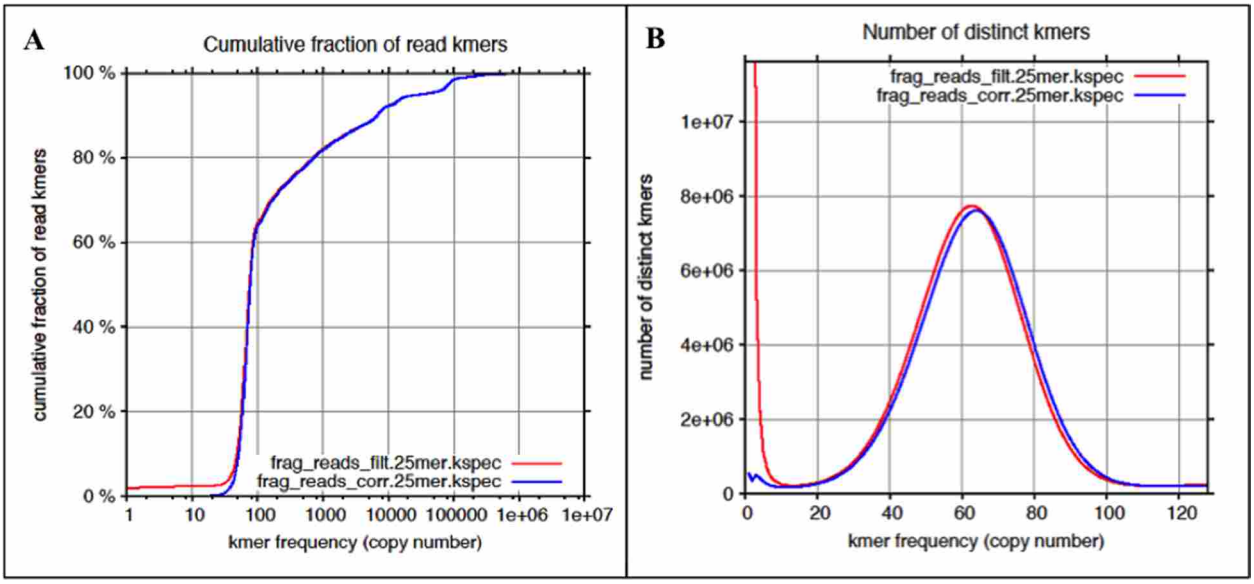


Figure 1. K-mer spectrum analysis of ALLPATHS-LG assembly of input reads. A) Cumulative fraction of kmers in the reads plotted as the frequency of the kmer. The spectrum for the filtered reads (red line) and corrected reads (blue line) begins at about 2% indicating that 2% of kmers in the reads have very low frequency and are most likely sequencing errors. Higher frequency kmers are associated with repetitive elements in the genome. The knee point at 64% indicates that at least 36% of the assembled genome is repetitive. B) Number of distinct kmers in the reads plotted as the frequency of the kmer. After the reads were corrected the low frequency spectrum is very close to zero indicating that error correction algorithm worked. The spectrum indicates no sequencing bias or evidence of large-scale heterozygosity in the genome.

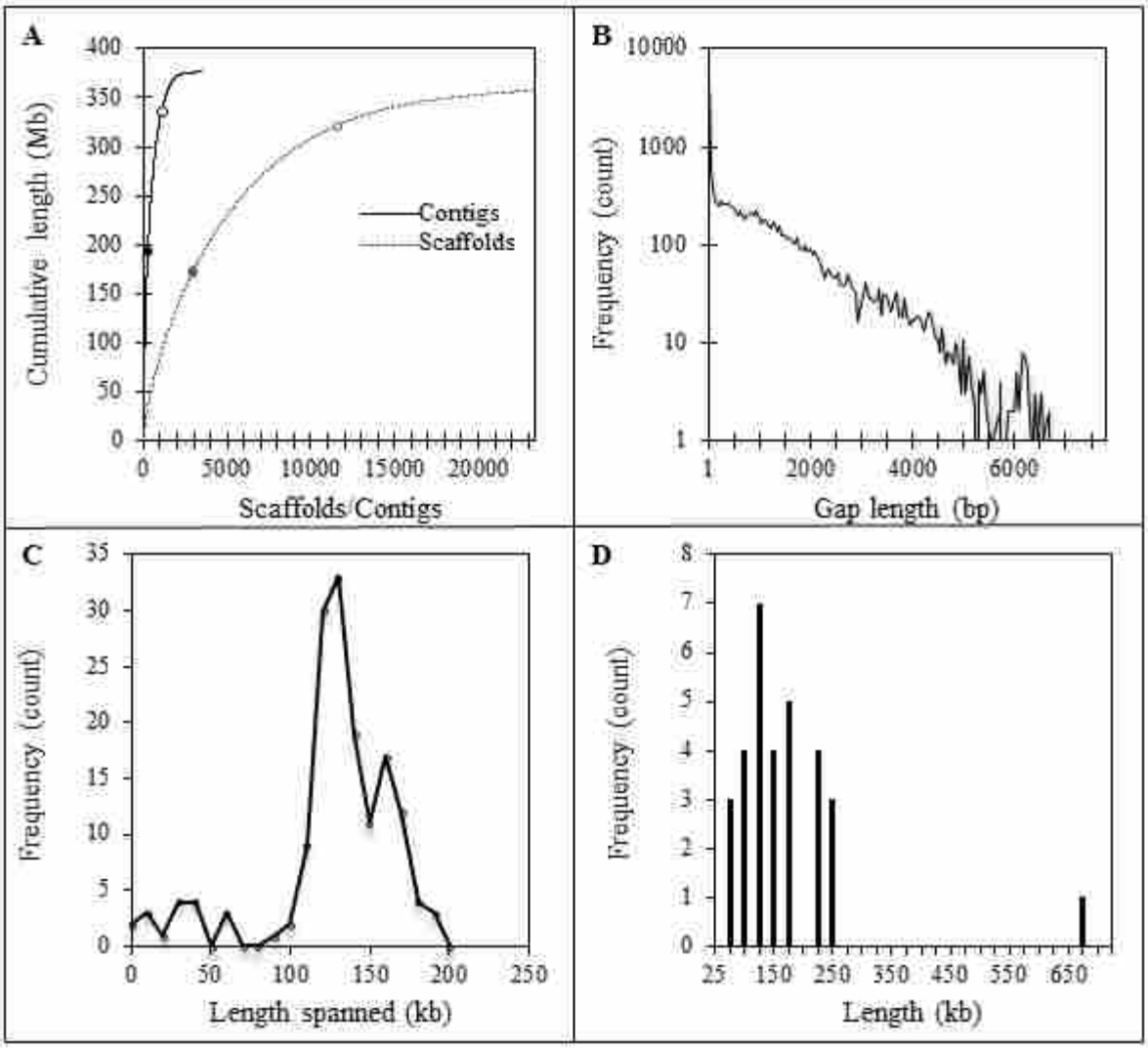


Figure 2. Results of the GapCloser V1.0 Assembly. A) Contigs and scaffolds were sorted by size in descending order. Cumulative length of the assembly was plotted by scaffold or contig. The N_{50} (filled in circle) and N_{90} (open circle) of the assembly are also shown. B) Distribution of gap length within the GapCloser V1.0 assembly. C) Distribution of the length of intervening sequence in scaffold/BES alignments. The distance for 158 paired BES that aligned to the same scaffold was plotted as a frequency distribution for each of the pairs. D) Distance between SNP pairs that aligned to the same scaffold were plotted as a histogram with 25,000 bp bins.

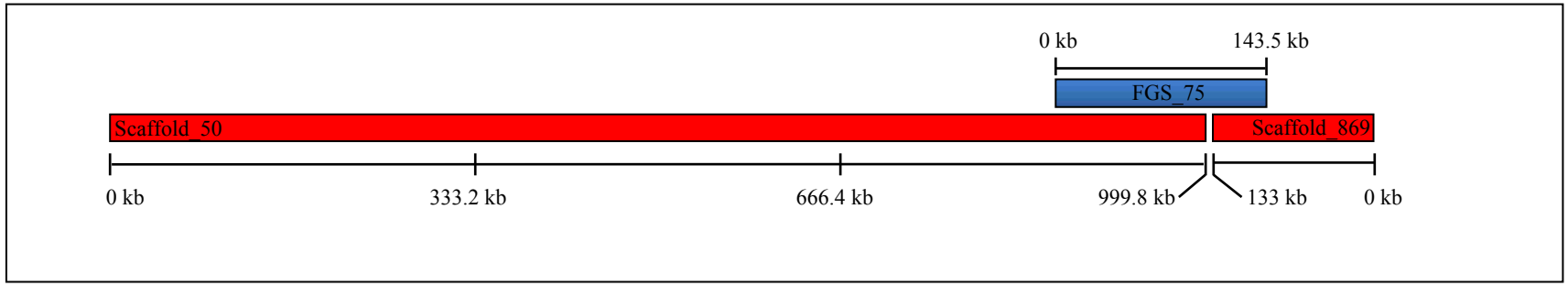


Figure 3. Example of alignments between the FGS BAC assembly and GapCloser V1.0 assembly. FGS scaffold_75, 143.5 kb in length, aligned to the last 97 kb of the GapCloser V1.0 scaffold_00050 with >99% identity and to the first 42 kb of the GapCloser V1.0 scaffold_00869 with >99% identity. Alignments were generated using MULAN (Loots and Ovcharenko, 2008).

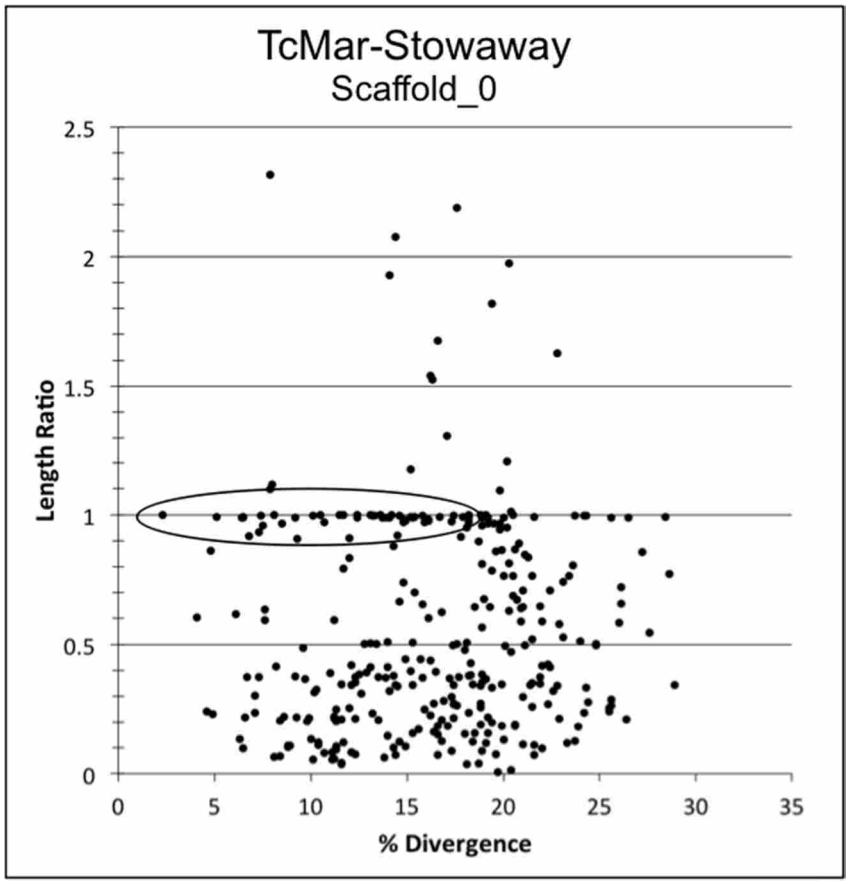


Figure 4. Length ratio by percent divergence plotted for 425 TcMar-Stowaway repetitive elements for scaffold_0 (2.5 Mb). Length ratio >0.9 and percent divergence <20 are considered potentially active in the genome (circled area). For ease of visualization, data is only presented graphically for scaffold_0.

Figure 5. Evidence of transposition in grain amaranths as seen by the areas of purple and green flowers on the inflorescence. Photo courtesy of Luz Gomez-Pando.

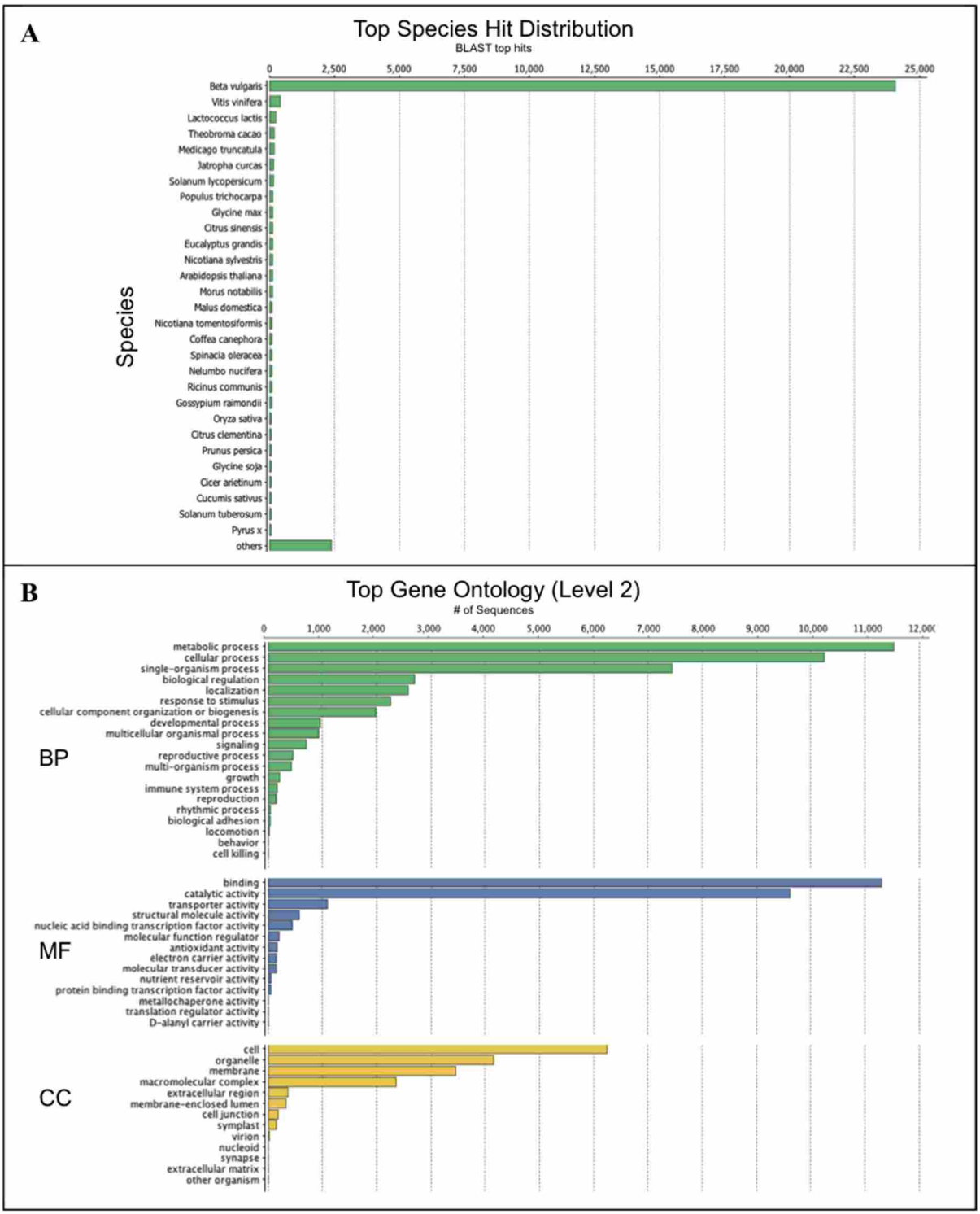
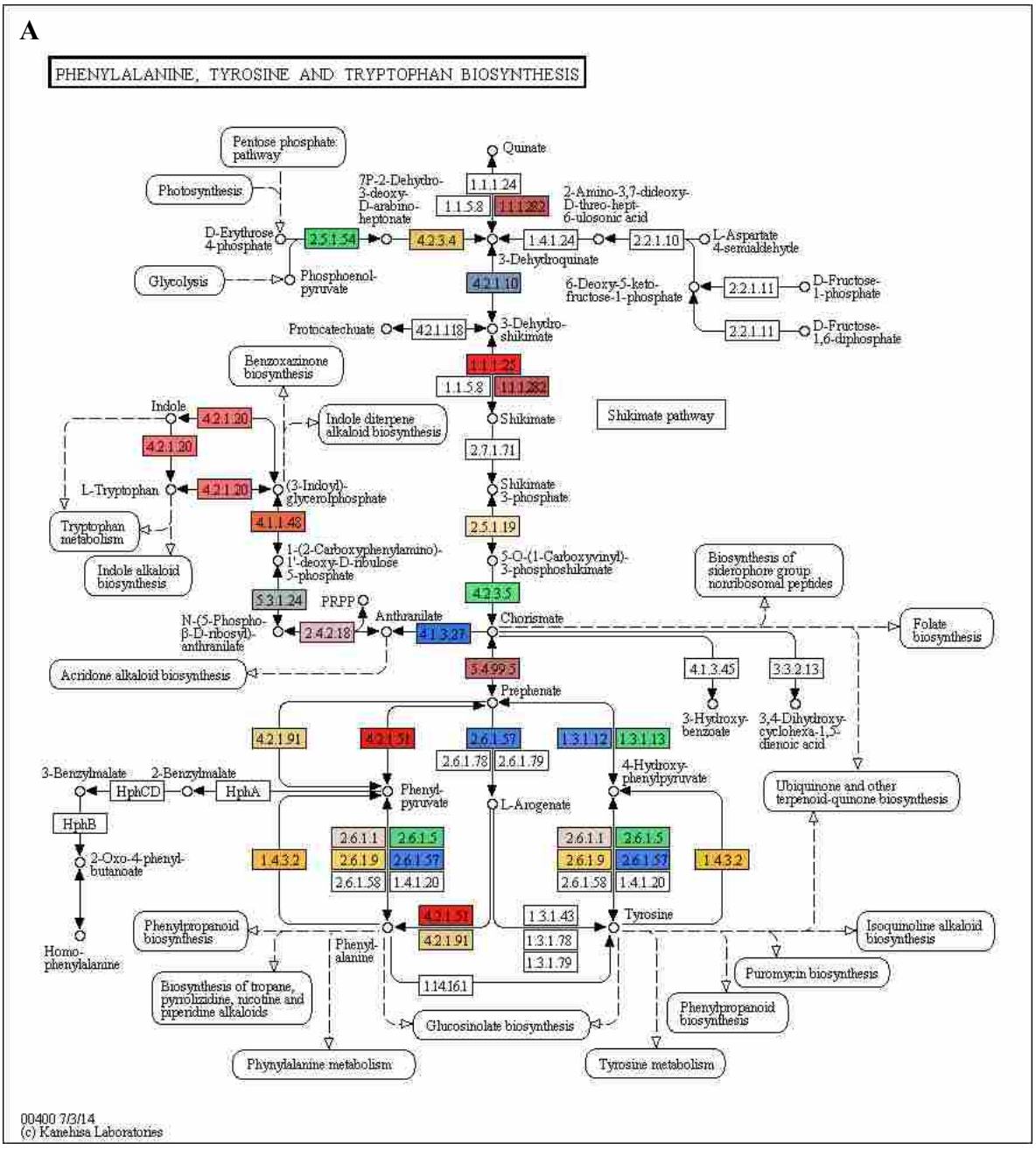
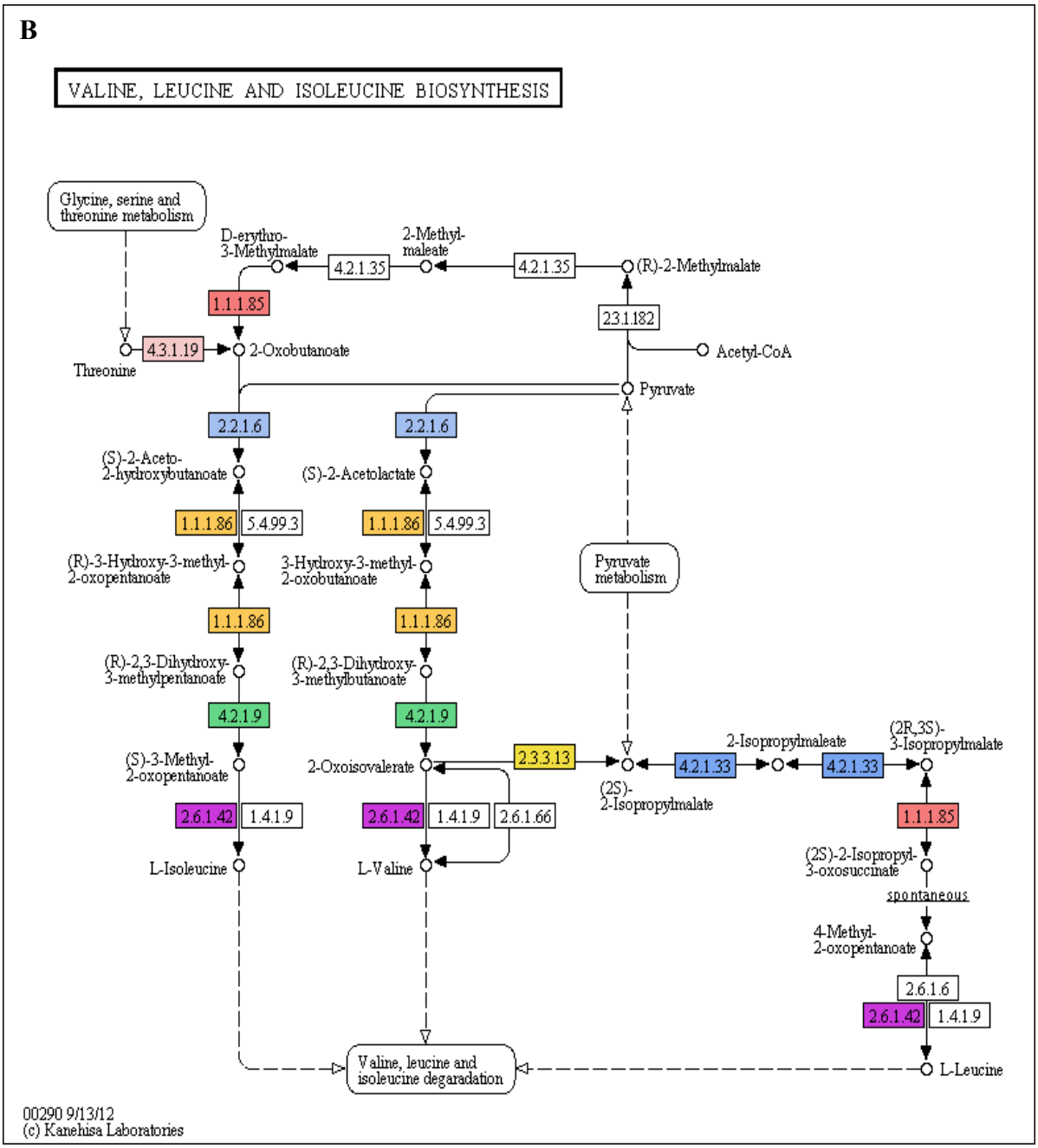


Figure 6. BLASTx protein matches and Blast2Go annotation of 65,947 transcriptome scaffolds. A) Top species alignments to amaranth sequences using the NCBI nr protein database. B) The most dominant GO terms (level 2) for all three B2G categories, biological process (BP), molecular function (MF), and cellular component (CC).

A


PHENYLALANINE, TYROSINE AND TRYPTOPHAN BIOSYNTHESIS
Pentose phosphate pathway
Photosynthesis
Glycolysis
D-Erythrose 4-phosphate
Phosphoenol-pyruvate
7P-2-Dehydro-3-deoxy-D-arabino-heptonate
2-Amino-3,7-dideoxy-D-threo-hept-6-ulosonic acid
Quinate
3-Dehydroquinate
L-Aspartate 4-semialdehyde
6-Deoxy-5-keto-fructose-1-phosphate
D-Fructose-1-phosphate
D-Fructose-1,6-diphosphate
Protocatechuate
3-Dehydro-shikimate
Benzoxazinone biosynthesis
Shikimate
Shikimate pathway
Indole
Indole diterpene alkaloid biosynthesis
L-Tryptophan
(3-Indoyl)-glycerolphosphate
Shikimate 3-phosphate
Tryptophan metabolism
Indole alkaloid biosynthesis
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose 5-phosphate
5-O-(1-Carboxyvinyl)-3-phosphoshikimate
Biosynthesis of siderophore group nonribosomal peptides
PRPP
N-(5-Phospho-β-D-ribosyl)-anthranilate
Anthranilate
Chorismate
Folate biosynthesis
Acridone alkaloid biosynthesis
Prephenate
3-Hydroxy-benzoate
3,4-Dihydroxy-cyclohexa-1,5-dienoic acid
3-Benzylmalate
2-Benzylmalate
HphCD
HphA
HphB
2-Oxo-4-phenyl-butanoate
Homo-phenylalanine
Phenyl-pyruvate
L-Arogenate
4-Hydroxy-phenylpyruvate
Ubiquinone and other terpenoid-quinone biosynthesis
Phenylpropanoid biosynthesis
Phenyl-alanine
Tyrosine
Isoquinoline alkaloid biosynthesis
Puromycin biosynthesis
Biosynthesis of tropane, pyrrolizidine, nicotine and piperidine alkaloids
Glucosinolate biosynthesis
Phenylpropanoid biosynthesis
Phynylalanine metabolism
Tyrosine metabolism
00400 7/3/14
(c) Kanehisa Laboratories

B


VALINE, LEUCINE AND ISOLEUCINE BIOSYNTHESIS
Glycine, serine and threonine metabolism
D-erythro-3-Methylmalate
4.2.1.35
2-Methyl-maleate
4.2.1.35
(R)-2-Methylmalate
1.1.1.85
23.1.182
Acetyl-CoA
4.3.1.19
Threonine
2-Oxobutanoate
Pyruvate
2.2.1.6
2.2.1.6
(S)-2-Aceto-2-hydroxybutanoate
(S)-2-Acetolactate
1.1.1.86
5.4.99.3
1.1.1.86
5.4.99.3
(R)-3-Hydroxy-3-methyl-2-oxopentanoate
3-Hydroxy-3-methyl-2-oxobutanoate
Pyruvate metabolism
1.1.1.86
1.1.1.86
(R)-2,3-Dihydroxy-3-methylpentanoate
(R)-2,3-Dihydroxy-3-methylbutanoate
4.2.1.9
4.2.1.9
(S)-3-Methyl-2-oxopentanoate
2-Oxoisovalerate
2.3.3.13
(2S)-2-Isopropylmalate
4.2.1.33
2-Isopropylmaleate
4.2.1.33
(2R,3S)-3-Isopropylmalate
2.6.1.42
1.4.1.9
2.6.1.42
1.4.1.9
2.6.1.66
1.1.1.85
L-Isoleucine
L-Valine
(2S)-2-Isopropyl-3-oxosuccinate
spontaneous
4-Methyl-2-oxopentanoate
2.6.1.6
2.6.1.42
1.4.1.9
Valine, leucine and isoleucine degradation
L-Leucine
00290 9/13/12
(c) Kanehisa Laboratories

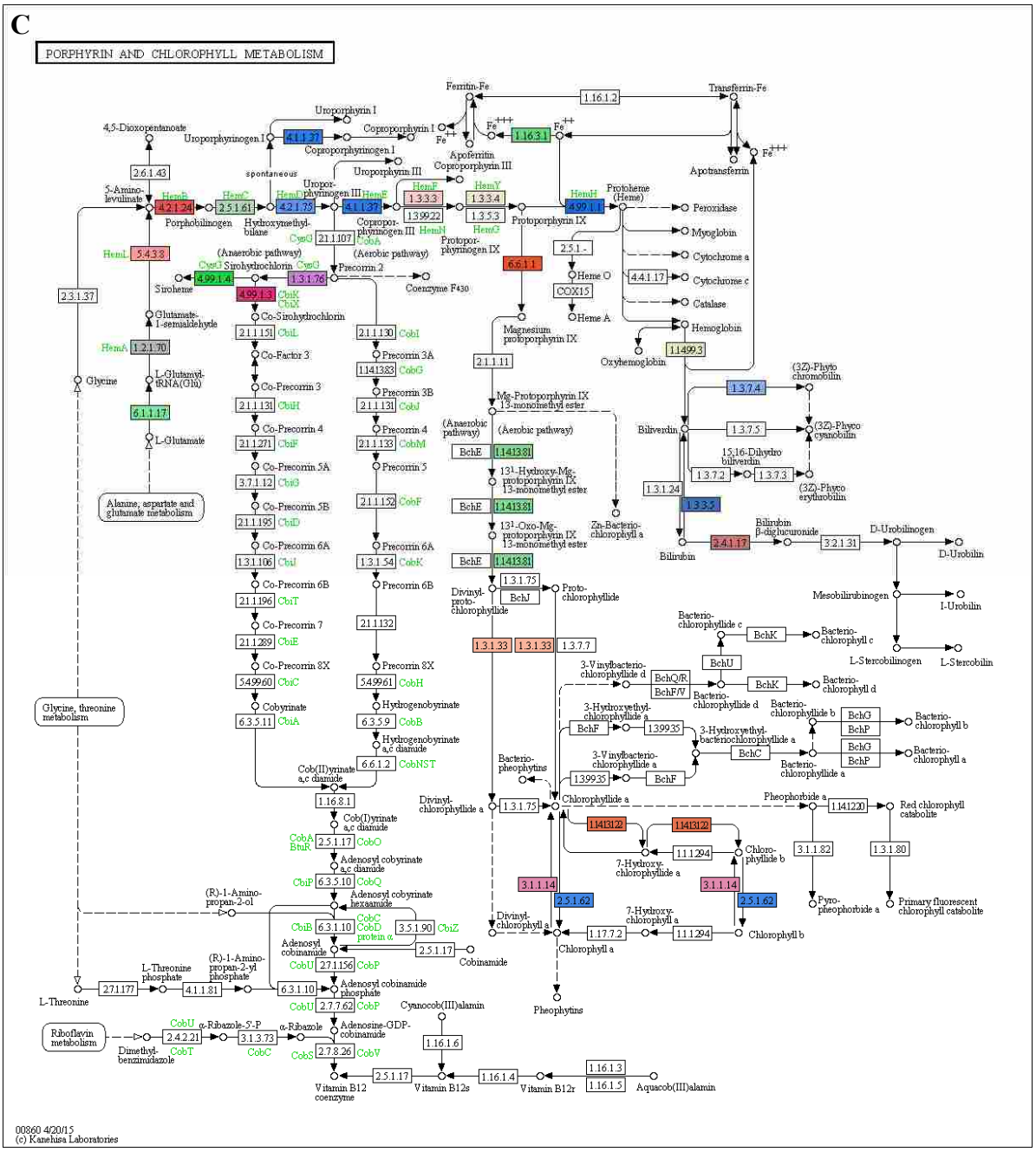


Figure 7. KEGG pathway maps of herbicide targets. Enzymes (EC #) identified in the amaranth transcriptome are shown in color. A) Biosynthesis of aromatic amino acids (phenylalanine, tyrosine, and tryptophan; map00400), B) Branched amino acids (valine, leucine, and isoleucine ; map00290) and C) Porphyrin and chlorophyll metabolism (map00860).

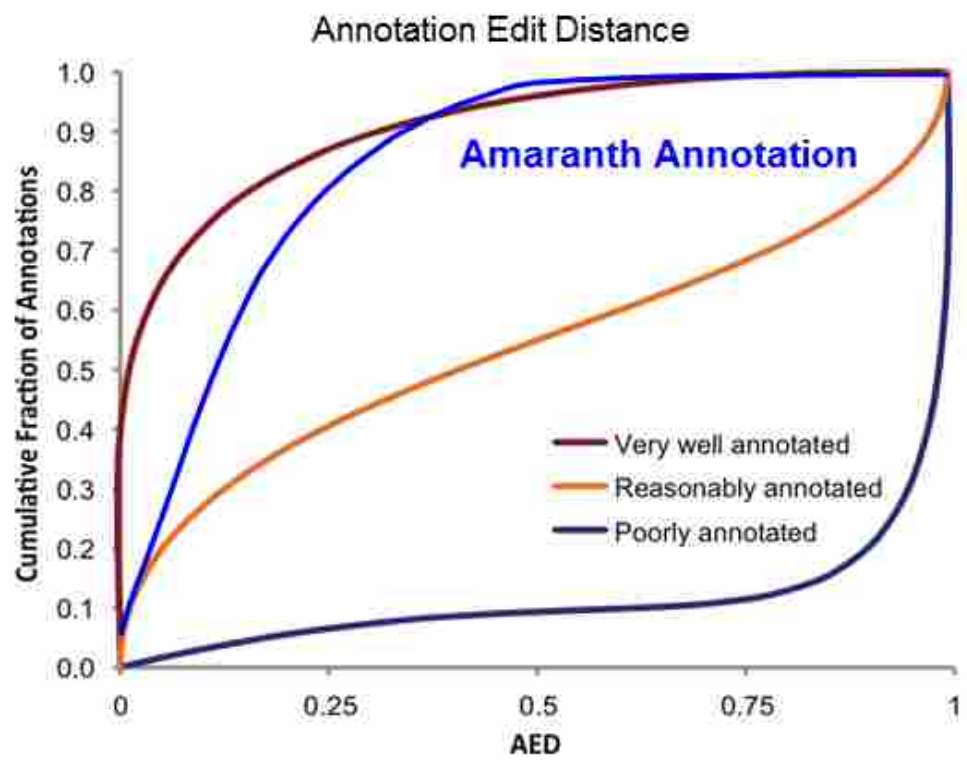


Figure 8. Quality assessment of the annotations generated by MAKER as determined by Annotation Edit Distance (AED). For illustrative purposes, the amaranth annotation quality (blue line), is shown with theoretical expectations for poor, reasonable and well annotated genomes. Greater than 80% of the amaranth annotations had gene predictions with an AED <0.25.

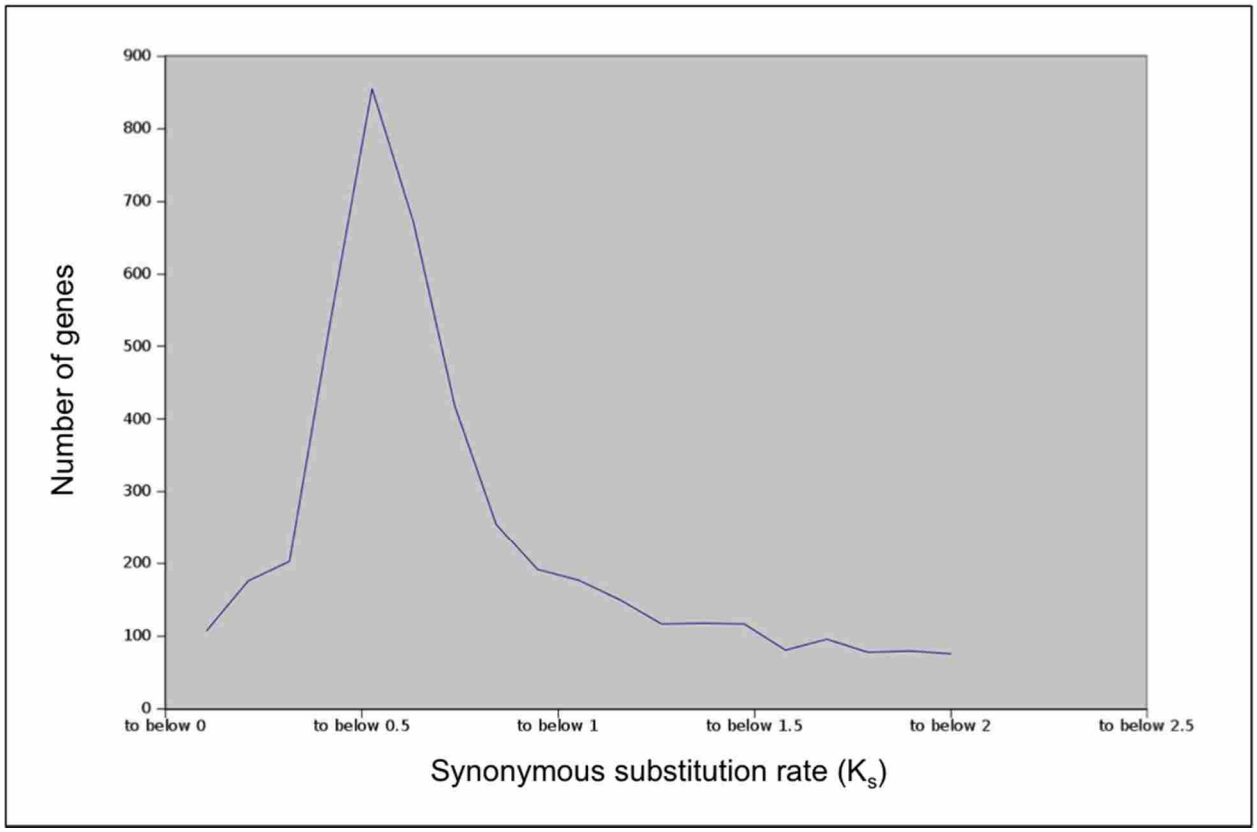


Figure 9. Rate of synonymous substitutions per synonymous sites (K_s) within 860 duplicated gene pairs as calculated by DupPipe. The peak in the graph indicates that a large number of gene pairs accumulated mutations at the same frequency indicating that they most likely originated at the same time through a WGD event.

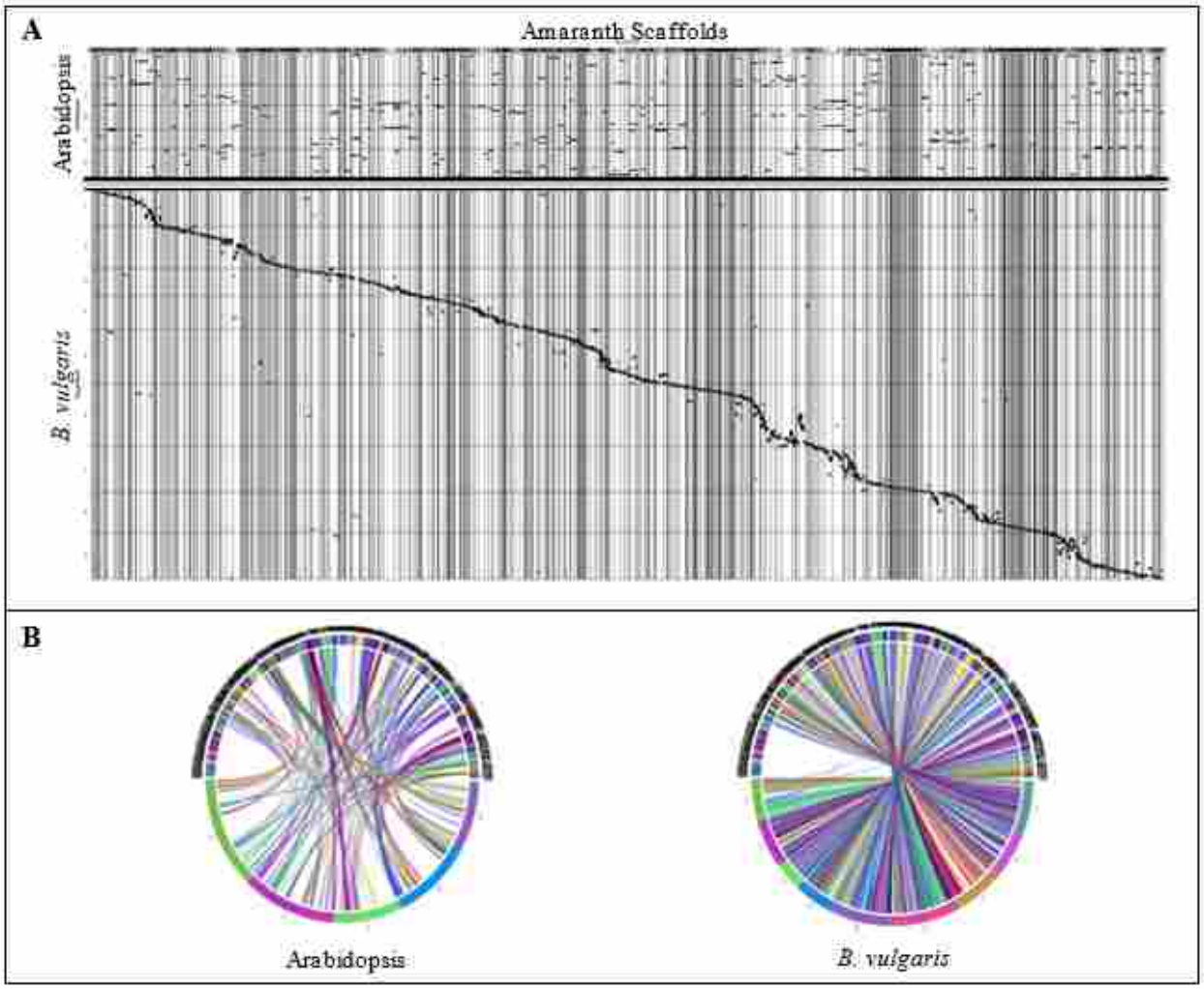


Figure 10. Syntenic relationship between *A. hypochondriacus, B. vulgaris* and *A. thaliana*. A) All GapCloser V1.0 scaffolds greater than 100 kb (943 scaffolds; x-axis) were aligned against the nine sugar beet chromosomes and five *Arabidopsis* chromosomes (y-axis). B) Circular display showing the synteny blocks as colored ribbons between the amaranth scaffolds and Arabidopsis or *B. vulgaris* chromosomes.

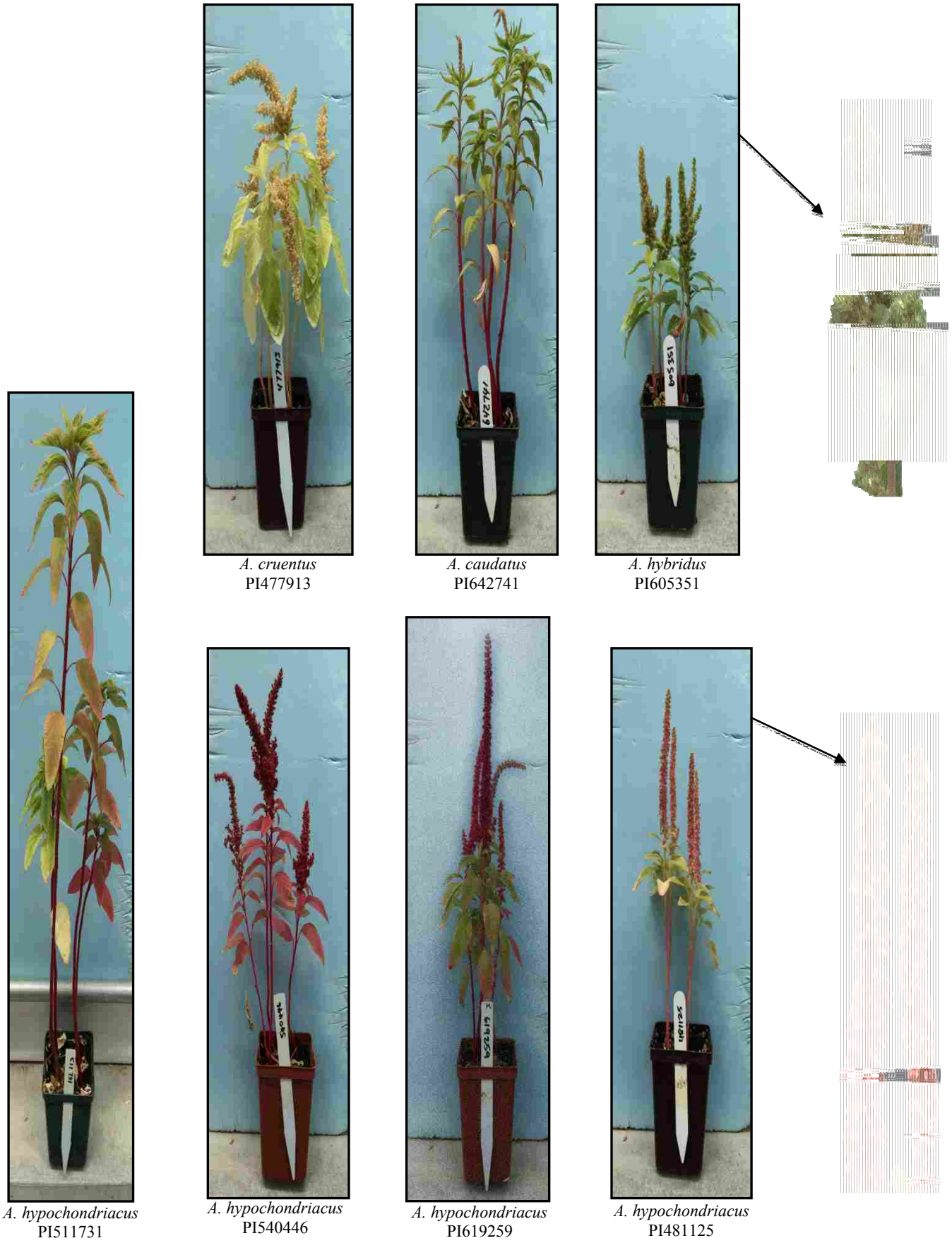


Figure 11. Images of greenhouse grown plants of the seven re-sequenced accessions. The inflorescences of PI605351 and PI481125 are highlighted.

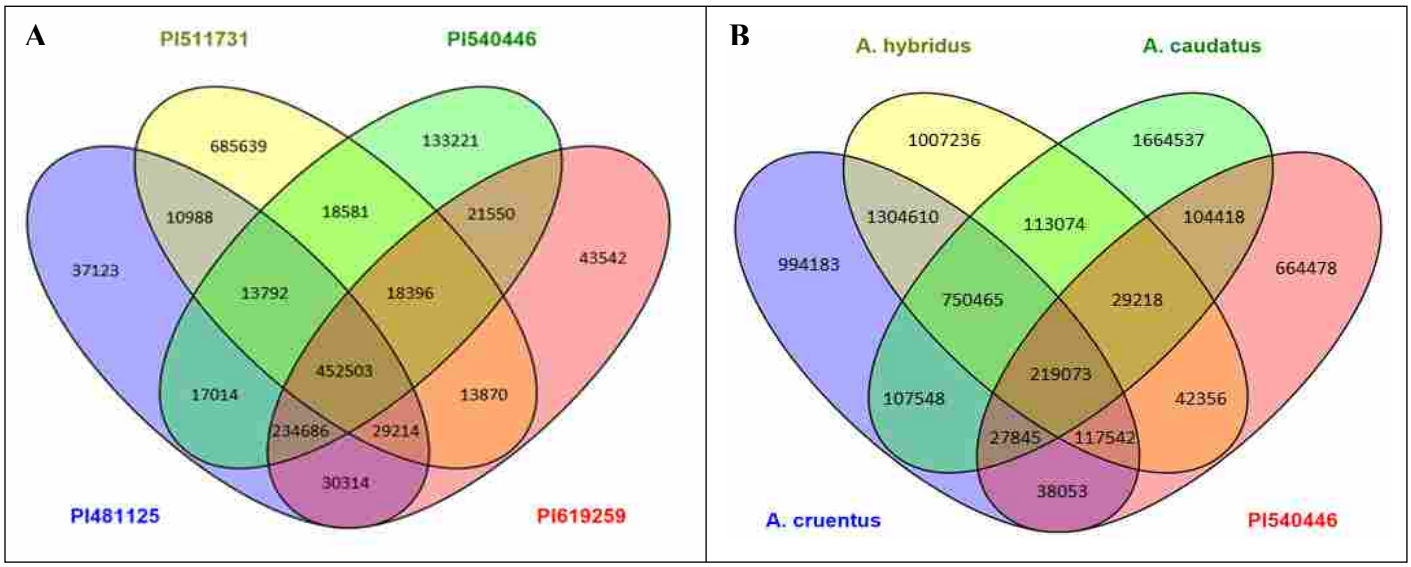


Figure 12. Comparison of SNPs identified by mapping reads of seven *Amaranthus* accessions to the GapCloser V1.0 assembly. The number of identical SNPs between accessions are shown. A) Venn diagram of the intraspecific SNPs identified in the *A. hypochondriacus* accessions. B) Venn diagram of the interspecific SNPs identified among *A. hybridus, A. cruentus,* and *A. caudatus*.

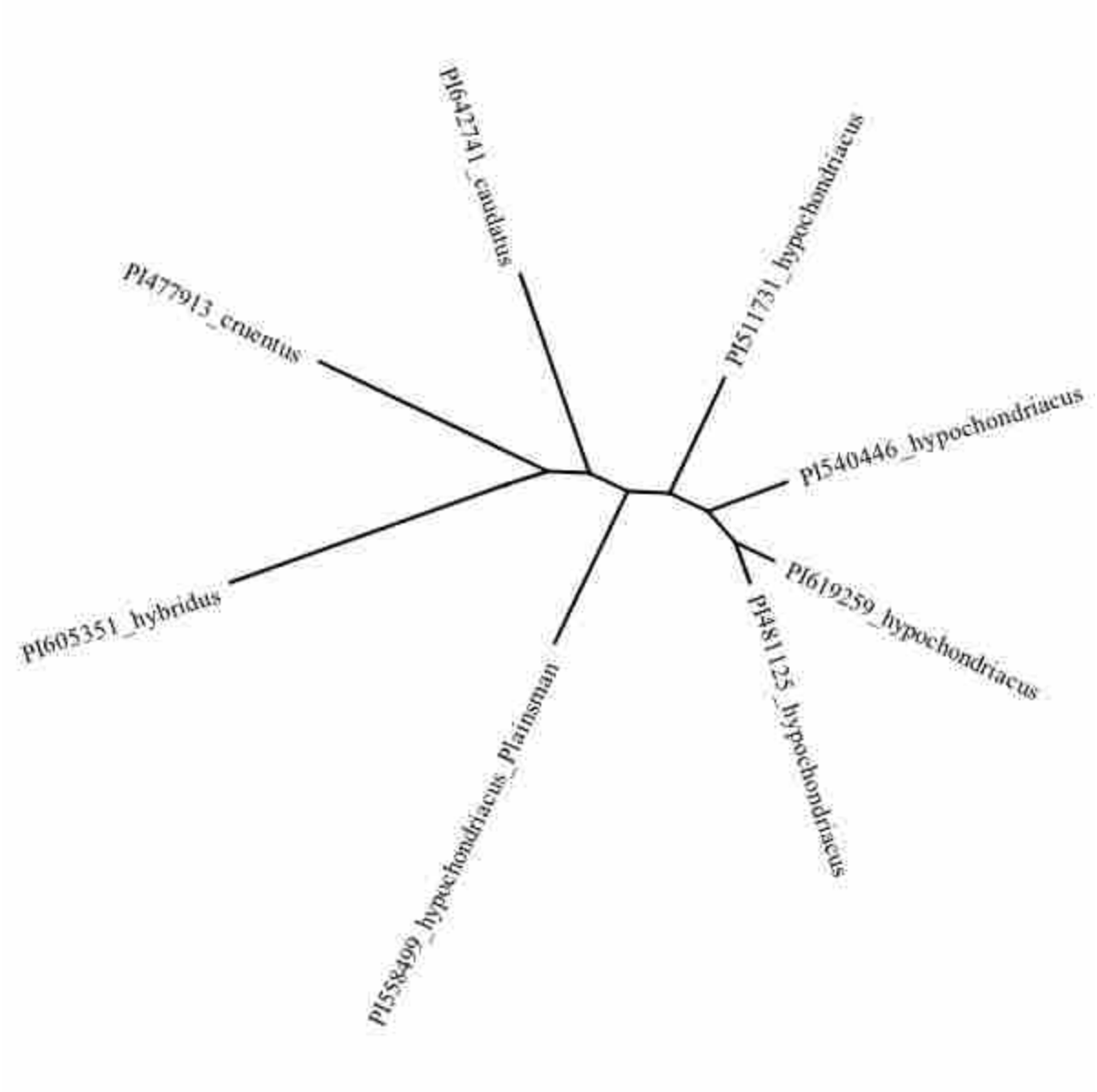


Figure 13. Unrooted neighbor joined tree showing the relationship of the seven re-sequenced accessions of amaranth based on all 7,495,570 identified single nucleotide polymorphisms. Bootstrap support values were 100% at each node. The *A. hypochondriacus* accessions form a distinct clade which includes PI481125, formerly classified as *A. caudatus*.

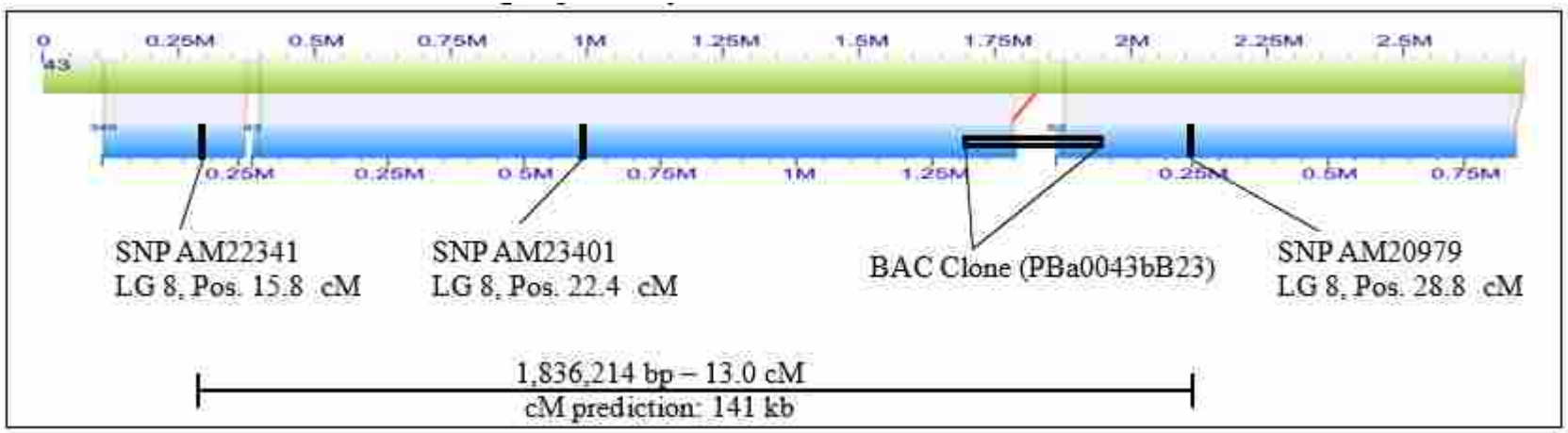


Figure 14. Hybrid assembly verification. Hybrid scaffolds are shown in green and GapCloser V1.0 assembly scaffolds are shown in blue. Linkage map information (SNPs) from linkage group 8 bridged three GapCloser V1.0 scaffolds (scaffolds 00027, 00071 00378) together into hybrid scaffold_00043. SNP AM22341 mapped to scaffold_00378 while SNPs AM23401 and AM20979 mapped to scaffolds 00027 and 00071 respectively. The three SNPs spanned a total of 13.0 cM and 1.8 Mb with a predicted cM length of 141 kb. SNPs AM22341 and AM23401 were separated by 6.6 cM while AM23401 and AM20979 were separated by 6.4 cM. BAC end sequences from BAC clone PBa0043bB23 spanned the gap between GapCloser V1.0 scaffolds 00027 and 00071. The total length spanned by the BAC was 216 kb.

CHAPTER 2: Literature Review

INTRODUCTION

The word amaranth is derived from ancient Greek meaning "everlasting" (Kauffman and Weber, 1990). It is a fitting name for a crop that once thrived in Pre-conquest America before its quick decline post-colonization, but has found new life in recent years. Grain amaranths have garnered increased interest because of their nutritional quality and tolerance to abiotic stress. This increased interest has led to a growth in amaranth research, available genetic tools, and food use worldwide.

Grain amaranths have traditionally been grown in underdeveloped or impoverished areas of the world. Improvement of breeding, harvesting, and nutritional characteristics could be stimulatory to the economies of these areas as well as provide a valuable food source where other crops have previously failed. Amaranth could also become a more important food crop globally as the world's population is expected to rise above 9 billion by 2050 and as traditional farming lands are becoming more marginalized by urbanization (Munns, 2002).

Many valuable genetic resources have been developed for amaranth, including RAPDs (Chan and Sun, 1997), SNPs (Maughan et al., 2009), linkage map (Maughan et al, 2011), BAC library (Maughan et al., 2008) and SSRs (Mallory et al., 2008), but a high quality reference genome has yet to be developed. Reference genomes are invaluable resources that greatly enhance the ability of researches to elucidate gene function, develop effective breeding programs, and other molecular processes. Here, I provide a brief literature review of grain amaranths and an introduction into plant genome sequencing and assembly.

GRAIN AMARANTHS

HISTORY

Ancient – Amaranth was one of the original crops cultivated on the American continent. Amaranths were first cultivated by the indigenous people of South and Central America over 8,000 years ago (Rastogi and Shukla, 2013). The earliest archaeological findings identified a cache of amaranth seed of *A. cruentus* in Tehuacan, Puebla, Mexico dated back to 4000 BCE (Iturbide and Gispert 1994). It is likely that *A. cruentus* migrated to this region from South America rather than being locally domesticated (Sauer, 1993).

Pre-Columbian –Various indigenous people throughout North and South Americas cultivated the grain amaranths (*Amaranthus hypochondriacus, Amaranthus cruentus, Amaranthus caudatus*) for the seed that played a vital role in the diet, economics, and culture of their societies (Sauer 1950, 1967). It was one of the four major crops used by the Aztecs. The ancient Aztecs used the *huauhtli* seeds, the native name for amaranth and chenopod grain, in a variety of ways (Sauer 1950). First, it was an integral part of the diet of indigenous people. The amaranth seed itself was nutritionally used in a variety of ways. One way the seeds were used was by popping them and then grinding them into flour that would be later used for tortillas and tamales (Iturbide and Gispert 1994). The seeds were also used to make a drink called *atole*, mixed with honey to make confections, and occasionally crushed and mixed with water to form a mush or gruel (Sauer 1950). Second, amaranth grain played a major role in the economy of the Aztec Empire. A tribute of 200,000 bushels of amaranth was required to be paid to Emperor Montezuma, the leader of the Aztec Empire at the time of the Spanish Conquest, from the 17 provinces that made

up the empire. This was compared to 230,000 bushels of beans and 280,000 bushels of maize (Sauer 1993). Lastly, amaranth was of central cultural importance to the Aztec people. At various times throughout the year, festivals would be held to honor one of the deities of the polytheistic Aztec people. One example was an annual festival held in May to honor the god of war, Huitzilopochtli. Amaranth flour would be mixed with maguey honey and form a paste called *tzoalli*. *Tzoalli* would be molded to represent whatever deity was being honored, in this case, Huitzilopochtli. The sculpted *tzoalli* would then be a focal point of worship and ritualistic sacrifice. Once all the rituals were completed, the idol would then be broken up and fed to the worshippers. It represented the bones and flesh of the worshipped god and was eaten in remembrance of the god's greatness. (Iturbide and Gispert 1994, Sauer 1967, 1993).

Decline in Use – After the arrival and colonization of the New World by the Spaniards in the early sixteenth century, cultivation of the grain amaranth in the Americas began to decline. Iturbide and Gispert (1994) outline three main reasons for the decline. First, different crops from the Old World were transported by the Spaniards to the New World and replaced some of the traditional crops grown by the indigenous people. Second, a loss of appreciation or distaste of the flavor of the grain contributed to its decline. Lastly, and possibly most notably, religion played a major role in the decline of amaranth use. Catholic priests who arrived from Spain observed the Aztec rituals and their use of amaranth and interpreted them to be perversions of the Holy Eucharist (Sauer 1950, Iturbide and Gispert 1994). While forcible destruction of amaranth crops never occurred, the cultivation and consumption of amaranth grain was greatly persecuted (Iturbide and Gispert 1994, Sauer 1993). The production of amaranth began to dwindle, and although production of amaranth persisted until the late 18th century, by the end of the 19th century amaranth was not listed as a grain crop in Mexico (Sauer 1950).

Amaranth Outside of the Americas – After the Spanish conquered most of what is today Latin America, amaranth use and cultivation nearly disappeared in this region. While amaranth cultivation was declining in the New World, amaranth was being introduced to the Old World. All three amaranth grain species were grown in the Old World but mainly as ornamentals. It is at this time that *A. hypochondriacus* and *A. caudatus* received their popular names of Prince-of-Wales feather and love-lies-bleeding, respectively (Sauer, 1967). *A. hypochondriacus* gained popularity in India and Sri Lanka during the 18th century as a grain crop and its cultivation spread into China and southern Siberia. *A. caudatus* was often grown alongside *A. hypochondriacus* in these regions as well. The grain of *A. hypochondriacus* also became an integral part of Hindu festivals and fast days. Its grain is the only food permitted to be eaten on fast days and certain festivals. The ritualistic importance of amaranth eventually led to it being introduced into East Africa during World War II. Here it was grown by local Indian communities and now hundreds of hectares of it are grown in Kenya annually (Sauer, 1993). Amaranth was also known to have been grown in other parts of Africa (Uganda, Sudan, and Angola) and in Central Asia (Afghanistan and Iran) during the 18th and 19th centuries (Sauer, 1967).

Present Day – Research on amaranth suggests that it could possibly be grown on marginal farming lands and provide nutrition to malnourished people around the globe. This once tropical plant can be grown at higher latitudes and can be harvested with wheat combines. In Central America amaranth seed is now popped and mixed with maguey honey to make a confection called *alegria* which is very similar to *tzoalli* made by the ancient Aztecs. Amaranth germplasm is cataloged and stored in at least 11 countries worldwide. The Rodale Research Institute has created a germplasm collection of amaranth with over 1400 accessions from 12 different species of amaranth that can be used for research purposes (Kauffman and Weber, 1990). The

germplasm collection in the USDA GRIN system has been collected from over 40 countries and consists of over 3300 accessions from over 40 species within the genus. More than 1500 of the USDA GRIN accessions are *A. hypochondriacus*.

ORIGIN OF GRAIN AMARANTHS

Evolutionary History - Grain amaranths have arisen through the domestication of weedy amaranth species. Two hypotheses have historically been proposed for how grain amaranths evolved from their weedy progenitors. The first states that all three grain amaranth species have their own individual progenitors, *A. caudatus* evolving from *A. quitensis* in South America, while *A. cruentus* evolved from *A. hybridus* in Central America and *A. hypochondriacus* from *A. powellii* in Mexico. The second theory states that all three grain species originated from *A. hybridus* (Sauer 1967). Recent studies using genetic markers have shown that *A. hybridus* appears to be the progenitor of all three grain amaranth species, giving convincing evidence to the second hypothesis (Mallory et al., 2008, Kietlinski et al., 2014). Kietlinski et al. (2014) also suggested that *A. hypochondriacus* and *A. caudatus* are more closely related to each other than they are to *A. cruentus.* This relationship is also seen in phylogenies created by Mallory et al. (2008) and Xu and Sun (2001). Two different domestication events may account for this relationship. One possibility is that *A. hypochondriacus* and *A. caudatus* may have originated from either a single domestication event in Mesoamerica or in the Andes. The other is that they originated from two separate domestication events of distinct but related *A. hybridus* (weedy amaranth) accessions in these two different locations. *A. cruentus* seems to have arisen from a separate domestication event that occurred in Guatemala or Central Mexico (Kietlinski, 2014). *A. caudatus* has repeatedly hybridized with weedy amaranth species, in particular *A. quitensis*, which complicates the elucidation of the origin of this species (Sauer, 1967, Kietlinski, 2014).

The evolutionary relationship of *A. hybridus* to the grain amaranths is supported by genetic analyses using single nucleotide polymorphism (SNPs) and single sequence repeats (SSRs) (Maughan et al., 2011, Mallory et al., 2008).

Plainsman Cultivar - In 1977 a cultivar of *A. hypochondriacus* was developed by the Rodale Research Center in Pennsylvania and was originally released in 1985 under the name K343 (Brenner, 1992). A Mexican landrace of *A. hypochondriacus* (PI 477917) was crossed with a Pakistani landrace of what was thought to be *A. hybridus* (PI 540446), but has subsequently been reclassified as *A. hypochondriacus*, to create this cultivar that was later renamed Plainsman (PI 558499). Earlier maturity, lighter seed color, and shorter plant height were all selected for when this cultivar was developed. Shorter plant height enables Plainsman to be more easily harvested using a combine. Plainsman matures in ~110 days making it one the earliest-maturing grain amaranth lines and allows it to be grown at higher latitudes where the growing season is shorter (Baltensperger, 1992). It was named after the high plains of the Midwestern United States where it was distributed by the University of Nebraska (Sauer, 1993).

BOTANICAL DESCRIPTION

Amaranth is an annual plant and one of the few C_4 dicotyledonous plants which can grow rapidly with little water loss via transpiration (Stallknecht and Schulz-Schaeffer, 1993, Sauer, 1993). Its dicotyledonous nature precludes amaranth from being classified as a true cereal, as true cereals are monocotyledonous grasses. Therefore, amaranth is referred to as a pseudo-cereal. The cultivated varieties of amaranth are monoecious with small unisexual flowers (Kauffman and Weber, 1990, Tapia, 1994).

Height, Leaves, and Inflorescences - The grains amaranths have a stem axis that leads to a large inflorescence. The three grain species of amaranth can range in height from 0.4m to 3.5m. Genetic improvements are being made to target plants that are 1.0m to 1.5m that will be better for combine harvesting. The inflorescences can be erect or decumbent and come in a variety of colors – green, yellow, orange, pink, red, purple, and brown (Tapia, 1994, Iturbide and Gispert, 1994, Stallknecht and Schulz-Schaeffer, 1993). Amaranth leaves are petiolate, oval, ovate-oblong, and lanceolate in shape and green or purple in color (Iturbide and Gispert, 1994 and Tapia 1994).

Seed - Grain amaranth seeds are round, smooth, slightly flattened and small, 0.9mm to 1.7mm in diameter. They are variable and light in weight with 1,000 to 3,000 seeds per gram (Stallknecht and Schulz-Schaeffer, 1993). Seed color varies from white, gold, pink, and black and the fruit is contained in a pyxidium (Tapia, 1994). Light and dark seeded varieties have been found in all three grain species of amaranth. Light seed color has been associated with a loss in dormancy and will germinate quickly after being planted. Amaranth seeds show epigeal germination. Seedlings appear three to four days after planting and a panicle begins to appear at two and a half months after being planted (Iturbide and Gispert, 1994). Prehistoric hunter-gatherers would harvest and eat dark seeded amaranth, but with the domestication of grain amaranths, light-colored seeds were selected because of their better taste and popping qualities. Light color in the seed is a recessive trait and is not seen in wild and weedy varieties of amaranth. Elimination of dark seeds form grain amaranths is very difficult since the domesticated varieties are easily cross pollinated with omnipresent *Amaranthus* weedy species (particularly *A. hybridus*). When amaranth was originally domesticated, it was likely selected for large seed head size, not the seed size; therefore the size of the seed itself has remained small.

TAXONOMY

Amaranthaceae family - Amaranth belongs to the family Amaranthaceae within the order Caryophyllales, which contains nearly 180 genera and 2,500 species. Herb and subshrub species make up the majority of the species in the Amaranthaceae family. Many genera in the family have been cultivated as ornamentals such as *Celosia* (cockscomb), *Gomphrena* (globe amaranth), and *Iresine* (bloodleaf) (Sauer, 1993). Other genera contain serious weeds and invasive species that pose a threat to economic growth and the environment including *Kali* (windwitch), *Alternanthera* (joyweed), and *Amaranthus* (amaranth). The grain amaranths species share the genus *Amaranthus* with some dangerous weedy species as previously mentioned. *Amaranthus* is not the only cultivated genera in the family. *Chenopodium* (quinoa and canahua), *Beta* (beet and sugar beet), and *Spinacia* (spinach) are all important genera within the family. Sugar beet (*Beta vulgaris*) being the most well studied and economically important of these plant species.

***Amaranthus* genus –** The genus of *Amaranthus* consists of approximately 60 species with most of them being native to the Americas and only 15 species native to the other four continents, excluding Antarctica. Amaranths, more commonly known as pigweeds, mainly grow on riverbanks, mountain and desert canyons, lakeshores, ocean beaches, and tidal marshes in disturbed or loamy soils with little competition from other plant species. The seeds are dispersed by water or birds who feed on them (Iturbide and Gispert, 1994, Sauer 1967).

Grain Species – Three *Amaranthus* species are considered grain crops, *A. hypochondriacus, A. cruentus,* and *A. caudatus*. They have previously been discussed and a continuation of this discussion will come hereafter, therefore no discussion on them will be made here.

Vegetable Species – Amaranths have been grown as a leafy vegetable in a wider range than that of grain amaranths. Vegetable amaranths are grown in environmentally diverse areas such as the Caribbean, Mediterranean, Islands of the South Pacific, southwestern United States, throughout Asia from Russia down to Southeast Asia, and into Africa. There are four primary species of amaranth that are cultivated as vegetable crops, *A. cruentus, A. tricolor, A. dubius,* and *A. lividus.* The young leaves of the other two grain species can also be used as a potherb (Stallknecht and Schulz-Schaeffer, 1993). Vegetable amaranths have been cultivated in Southeast Asia for over 2,000 years and were cultivated throughout Europe during the Middle Ages. Two other species of amaranth, *A. hybridus a*nd *A. palmeri,* were consumed by Native Americans in the deserts of North America during the summer months until corn and beans were able to be harvested (National Academy of the Sciences, 1984).

Weedy Species – Weedy amaranths are commonly known as pigweeds. Many of these weedy species are considered as some of the most serious weeds in the world. *A. retroflexus* (redroot pigweed) and *A. viridis* (slender amaranth) are two of the most widely distributed weeds in the world. *A. retroflexus* has shown to reduce corn crop yield up to 5% in field studies (Knezevic et al., 1994). *A. palmeri* is a very troublesome weed in cotton, soybean, and peanut fields in the southeastern United States. It has been shown in recent years that *A. palmeri* has gained resistance to the herbicide glyphosate which makes it even more problematic (Culpepper et al., 2006). Other weedy species include *A. hybridus* (smooth amaranth), the progenitor of the grain amaranths, and *A. spinosus* (spiny amaranth), both of which are ranked among the 18 most serious weeds worldwide. *A. spinosus* alone has been shown to be troublesome to various crops worldwide including, sugarcane, sorghum, pineapple, upland rice, and others (Chauhan and Johnson, 2009).

Other Species and Uses – All three of the grain amaranth species are grown both as ornamentals and as the vegetable species. *A. tricolor*, also known as Joseph's coat after the Biblical figure, is another ornamental amaranth species. Native Americans used amaranth, specifically *A. cruentus,* to create a reddish dye (Sauer, 1967). Two species, *A. pumilus* (seabeach amaranth) and *A. brownii*, in this genus are listed as threatened or endangered on the U.S. Fish and Wildlife Endangered Species Database.

ADAPTATIONS

Amaranth has been able to adapt in semiarid environments that are susceptible to both salinity and drought stress (Omami and Hammes, 2006). The plant does this by making osmotic adjustments without wilting or dying (Tucker, 1986). At low concentrations of salt (25mM), increased germination rates are observed in grain amaranths and this rate can be increased to even higher levels if the seeds are subjected to seed priming before they are planted (Moosavi et al., 2009, Macler and MacElroy, 1989). Amaranth's ability to grow in these conditions makes it an attractive crop to replace maize and traditional cereal crops in semiarid environments and marginal farming areas. Currently 7% of the world's land is affected by salinity and, with irrigation and urbanization, this number is expected to rise, thus increasing the importance of salt tolerant crops such as amaranth (Munns, 2002). Research has already shown that amaranth can quickly adapt to new environments and thrive in areas where conditions are not suited for the cultivation of traditional cereal crops (Gupta and Gudu, 1991). This makes amaranth a very attractive crop to attempt to be grown in underprivileged areas of the world with marginal farming lands that have been previously uncultivated due to the limitations of traditional grain crops.

NUTRITION

Starch – Starch is the main carbohydrate component of amaranth grain comprising 48 to 69% of the seed. On average the seed is 60% starch which is lower than other comparable grains such as wheat, corn, and rice which are 66%, 67%, and 75% starch respectively (Cai et al., 2004). Amaranth starch is stored in extremely small granules that measure 0.8 to 2.5µm in size located in the perisperm. They are spherical or polygonal in shape with a rare, almost crystalline, structure and are the smallest granules ever recorded (Saunders and Becker, 1984). Granules in other grains range for 3 to 8µ in rice and 15 to 100µm in potatoes (Arendt and Zannini, 2013). Amylose content in amaranth grain ranges from 0 to 17% depending on the genotype (Konishi et al., 1985). *A. hypochondriacus* and *A. cruentus* have, on average, 7.2% and 7.8% amylose content, respectively (Yanez, 1986 et al., Qian and Kuhn, 1999). Amylopectin makes up most of the starch content in amaranth ranging from 92 to 95%. Amaranth starch also has good stability when exposed to freezing and thawing with its low peak viscosity and stable viscosity when exposed to temperature change. Starch properties of amaranth make it attractive for use in instant soups, gravies, and dressings, as well as baked goods such as pastas, breakfast cereals, and muffins. Additionally non-food uses include dusting powder in cosmetics, biodegradable plastics, and laundry starch (Choi et al., 2004).

Proteins – Protein content in amaranth grain is 16 to 18%, which is over 50% greater than other cereal crops (Gupta and Gudu, 1991). According to the FAO/WHO Nutritionist's Protein Value chart, amaranth scored higher than any other grain (75 out of a possible 100) and even higher than cow's milk (72). Also, when amaranth flour is mixed with maize, it reaches the perfect score of 100 (Arendt and Zannini, 2013). Amaranth also received a score of 90.4% in the Essential Amino Acid Index which is comparable with an egg. High levels of the essential amino

acid lysine and other amino acids arginine, histidine, and threonine are seen in amaranth. Valine, leucine, and isoleucine are in lower levels in amaranth and seem to be limiting amino acids. This high protein level and amino acid profile make amaranth a useable protein substitute for a meal (Piskarikova et al., 2005).

Lipids – The oil content of amaranth seeds ranges from 1.9 to 8.7% with an average of 5%, which is higher than cereal grains such as maize, rice, and wheat which have oil content of 4.5%, 2.1%, and 2.1% respectively (Belitz and Grosch, 1999). Triacylglycerols make up 80.3 to 82.3% of amaranth oil with phospholipids accounting for 9.1 to 10.2% and squalene levels ranging from trace amounts to 8%. The fat content of grain amaranth is higher than in true cereals (Rastogi and Shukla, 2013). Three major fatty acids are seen in triacylglycerols of amaranth oil: palmitic (22.2%), oleic (29.1%), and linoleic (44.6%) (Gamel et al., 2007, He and Corke, 2003). The overall degree of unsaturation of fatty acids is 75% (Rastogi and Shukla, 2013). Amaranth oil also has been shown to have good oxidation stability which is greater than what is seen in sunflower oil (Gamel et al., 2007).

Squalene is a 30-carbon terpenoid and is a precursor of cholesterol biosynthesis (Amicarelli and Camaggio, 2012). Squalene has many applications in the cosmetic and skincare industry, as well as a lubricant for precision instruments (He and Corke, 2003). Squalene has also been shown to have numerous health benefits including a reduction in side-effects of cancer treatments and lowering of cholesterol (Reddy and Couvreur, 2009, Martirosyan et al., 2007). Squalene is commonly extracted from the livers of sharks, but concern about marine life limits this practice. Amaranth could be an economically viable source of squalene (Gamel et al., 2007).

Minerals – Mineral content of amaranth is twice as high as traditionally used cereals. Calcium, magnesium, and iron are particularly high, while phosphorus and potassium are also seen in elevated levels (Alvarez-Jubete, 2009). The high levels of iron in amaranth could be used to fight iron-deficient anemia (Caselato-Sousa and Amaya-Farfan, 2012). Amaranth flour can also be used to improve the quality of gluten-free foods for people living with celiac disease as current gluten-free products have poor nutritional value (Alvarez-Jubete, 2009).

Vitamins and other compounds – While amaranth is not a good source of vitamins, it does have elevated levels of vitamin E, vitamin B2 (Riboflavin), and vitamin C (ascorbic acid) (Gamel et al., 2006). Saponins are found in very low levels (0.1%) in grain amaranths, unlike in its pseudocereal relative quinoa, and are completely safe for human consumption (Oleszek et al., 1999). Phytic acid is also present in amaranth (0.3 to 0.6%) and is a reserve of phosphate. Phytate has been shown to interfere with mineral adsorption and could possibly lower cholesterol in humans (Arendt and Zannini, 2013, Rastogi and Shukla, 2013). Amaranth is considered a good source of insoluble fiber with a content of 4.2% (Caselato-Sousa and Amaya-Farfan, 2012). Both insoluble and soluble fibers have known health benefits such as reducing cholesterol and promoting gut health. Amaranth flours have been shown to have antioxidant activity due to flavonoids (polyphenols from secondary metabolites) found in the seed. Three flavonoids have been identified, rutin, isoquercitrin, and nicotiflorin, and several health benefits are known to be caused by these compounds (de la Rosa et al., 2009). Anti-nutrients are also present in amaranth grain, particularly oxalates that could be problematic for people with celiac disease. The intake of oxalates can result in calcium and magnesium deficiencies which then can lead to kidney stones (Gelinas and Seguin, 2007).

GENETIC STUDIES

As grain amaranths have garnered increased attention because of their nutritional characteristics, more genetic research has been conducted to understand this underutilized grain. Grain amaranths are considered to be paleo-allotetraploids. *A. hypochondriacus* and *A. caudatus* are 2n=32 while *A. cruentus* is 2n=34. The extra chromosomes in *A. cruentus* appeared after the polyploidization event by primary trisomy (Greizerstein and Poggio, 1994). *A. caudatus* has a putative genome size of ~500 Mb, while *A. hypochondriacus* and *A. cruentus* have reported genome sizes of ~466 Mb (Bennett and Smith, 1991).

Draft Genome – In 2014 Sunil et al. published a draft genome of *A. hypochondriacus*. The seeds that were used to grow the plants were obtained from farmers in northern Karnataka, India. Five libraries were made for genome sequencing, one paired-end library with an insert size of 300 bp and four mate-pair libraries with insert sizes of 1.75, 3, 5, and 10 kb. The generated libraries were assembled using the SOAPdenovo31mer assembler. There were 367,441 scaffolds created that had an assembled length of 645 Mb but consisted of 58% N's. The assembly had an N_{50} of 35,089 bp with 34% G+C content. A transcriptome was also assembled consisting of 136 Mb that were in 57,658 contigs with an estimated 24,829 proteins encoded. Significant homology was found between the genome assembly and a transcriptome published by Delano-Frier eat al. (2011) with 83.3% of the bases in the transcriptome aligning back to genomic scaffolds. Using GO analysis, greater than 80% of the putative genes have been deciphered. Synteny was shown in 76 of the 100 longest scaffolds to the *Beta vulgaris* genome which is the closest genome to amaranth that has been assembled. Another genomic feature, repeat content, was explored and showed that *A. hypochondriacus* had 13.76% repeat content, far below the 63% that is seen in *B. vulgaris* (Sunil et al., 2014).

Transcriptome – The first transcriptome of *A. hypochondriacus* was generated in 2011 by Delano-Frier et al. (2011). Using 2,700,168 454-sequence reads 21,207 (20,408 isotigs and 799 contigs) high quality sequences were assembled that ranged in size from 80 to 3,379 bp. After assembling the reads into contigs, 178,636 reads remained as singletons. A total of 5,113 clean singletons were identified after quality control measures were used on all 178,636 singletons. Approximately 82% of all sequences had significant hits when aligned to sequences in the nr database at NCBI. The raw sequence files are publicly available in the Sequence Read Archive at NCBI under the study number SRP006173 (Delano-Frier et al., 2011).

BAC Library – A bacterial artificial chromosome (BAC) library has been developed of the grain amaranth, *A. hypochondriacus* (Maughan, et al. 2008). This library consists of 36,864 clones with an average insert size of 147 kb that amounts to approximately 10.6x coverage of the genome. Of the 36,864 clones, less than 2% contained no insert and 93% contained inserts over 10 kb while fewer than 7% consisted of organellar DNA. BAC end sequences (BES) of 384 clones were obtained to determine if the library was of sufficient quality for whole genome sequencing. The sequenced clones produced 728 reads that amounted to 563 kb of high-quality sequence data with 34% of the BES having detectable homologs in the RefSeq database. Acetolactate synthase (ALS) and protoporphyrinogen oxidase (PPO) are two genes that are targets of herbicides. Amaranths have gained resistance to herbicides that target these genes (Heap, 1997, Heap, 2004). To test the utility of the BAC library, these two genes were probed for and sequenced successfully, thus proving that this library can be used to study genes of interest in the *Amaranthus* genus (Maughan, et al., 2008).

Mircosatellite Markers – Microsatellites or single sequence repeats (SSRs) are valuable genetic markers that can be used in population studies and for identification and comparison between

different species or accessions within a species. Microsatellite markers have been developed for the amaranth genus. These markers were originally designed for the grain amaranths, but have been shown to have utility amongst the weedy species as well, and have potential uses in the leafy and ornamental varieties. Microsatellites were first developed from 1,457 clones that were generated from three different microsatellite enriched libraries for the AAT, AAC, and AT repeats. Microsatellites were also obtained from BES as previously mentioned (Maughan et al., 2008). A total of 382 microsatellites were identified from these four sources. Of the 382 identified microsatellites, 179 were found to be polymorphic. From the polymorphic microsatellites, 731 alleles were identified ranging from two to eight alleles per locus with an average of four alleles per locus. Heterozygosity (H) among the polymorphic microsatellites had an average of 62% with 59 (33%) of the SSRs being highly polymorphic ($H \geq .7$). When observing the individual grain species, 129 of the 179 polymorphic microsatellites were polymorphic in *A. hypochondriacus*, while 123 in *A. cruentus* and 136 in *A. caudatus* were polymorphic. In the weedy species, 177 (>99%) of the 179 polymorphic microsatellites amplified in *A. hybridus* while 158 (88%) and 141 (78%) amplified in *A. retroflexus* and *A. powellii*, respectively, proving their utility amongst the weedy species of amaranth (Mallory et al., 2008).

Random Amplified Polymorphic DNAs (RAPDs) – RAPDs of amaranth were first generated by Transue et al. (1994). Analysis of 29 polymorphic fragments separated 29 known accessions into three distinct morphological groups. This analysis showed that the RAPDs could be used to identify unknown grain amaranths by species. Further studies on amaranth have been performed using RAPDs and also isozymes. RAPD polymorphism values have been shown to be significantly lower (39.9%) in grain amaranths than in leafy (51%) and wild amaranth species

(69.5%). Similar results have also been observed in isozyme data. The reduction in polymorphism of grain amaranths is likely due to genetic bottlenecks in the process of speciation or strong directional selection throughout the process of domestication. Higher levels of polymorphism are seen in the vegetable amaranth species, as compared to grain amaranths, because of a more recent domestication event or lack of selection pressure during domestication (Chan and Sun, 1997).

Single Nucleotide Polymorphism (SNP) Discovery – SNP discovery in amaranth was accomplished using four mapping populations that were created by crossing three *A. caudatus* lines (Ames15170, PI 553073, and PI 642741) and one *A. hypochondriacus* line (PI 481125). PI 481125 was classified as *A. caudatus* when this study was performed, but in a later study was reclassified to *A. hypochondriacus* (Maughan et al., 2011). These four populations were labeled, Pop1-3 (Ames15170 x PI 553073), Pop1-4 (Ames 15170 x PI 642741), Pop2-3 (PI 481125 x PI 553073), and Pop2-4 (PI 481125 x PI 642741). A genomic reduction using restriction enzymes was performed to remove repetitive sequence and maximize the number of identified SNPs. A total of 27,658 SNPs were identified across all four populations with a high of 11,047 SNPs identified in Pop2-3 and only 140 SNPs found in Pop1-3. SNP density was observed to be 1/98,915 in Pop1-3, 1/2,748 in Pop1-4, 1/1,389 in Pop2-3, and 1/1,457 in Pop2-4. The average base coverage of the identified SNPs was 20X and all SNPs had a minor allele frequency from 30 to 50%. Validation of 35 SNPs from Pop2-4 was also performed. Re-sequencing of the 35 SNPs revealed that 34 (97%) were validated as expected. All identified SNP sequences can be found in GenBank in dbSNP under accession numbers ss161123993 to ss161151650 (Maughan et al., 2009). The utility of SNPs in amaranth research has been shown to be useful in assessing diversity of different species (Jimenez et al., 2013) and as a gene marker (Park et al., 2011).

Linkage Map Development – An interspecific mapping population, Pop2-4, used by Maughan et al. (2009) for SNP discovery, was utilized for the development of a linkage map in amaranth. In the Pop2-4, 11,038 SNPs were identified. From these putative SNPs, 419 SNP assays were designed using KASPar chemistry and screened against a diversity panel of grain and weedy amaranths on the Fluidigm nanofluidic 96.96 dynamic array system. KASPar chemistry uses competitive binding of allele-specific primers for PCR amplification and subsequent genotypic calling. In this chemistry two forward primers are used in the PCR reaction that differ at a single nucleotide, the putative SNP, and the reverse primer is common to both. A different fluorescent tag, FAM or VIC, is associated with each forward primer and depending on the fluorescence seen genotypic calls can be made. Of the 419 SNPs that were genotyped using KASPar chemistry, 411 produced definite genotypic clusters. Pairwise linkage analysis, with a minimum logarithm of the odds (LOD) scores of 5.0, was used to group all 411 SNPs in to 16 linkage groups that likely correspond to the 16 haploid chromosomes in amaranth. The linkage map spanned an estimated distance of 1288cM with an average distance of 3.1cM between SNPs. The vast majority (93%) of these SNPs were linked at a distance of less than 10cM (Maughan et al., 2011). The development of SNP markers and a linkage map provide valuable resources for further amaranth research in the development of better breeding programs and other genetic studies including the development of a high quality genome assembly.

Of the 419 SNPs screened against the grain amaranths and A. hybridus, 414 were shown to be polymorphic. The diversity panel of the three grain amaranths and the putative progenitor *A. hybridus*, revealed that *A. hypochondriacus* was the most polymorphic grain species while *A. cruentus* was the least. The 414 polymorphic SNPs were also screened against three other weedy species of amaranth to test their utility and transferability within the *Amaranthus* genus. In both

A. retroflexus and *A. powellii* 256 SNPs (62%) produced high confidence calls while 158 SNPs (38%) produced similar calls in *A. tuberculatus*. These results prove the utility and transferability of these SNPs within the *Amaranthus* genus.

GENOME ASSEMBLY

Genome Sequencing – *Arabidopsis thaliana* was the first plant genome sequenced (The Arabidopsis Genome Initiative, 2000). Between 2000 and 2009 only twelve additional plant genomes were sequenced due to the expensive process of sequencing and assembly using the Sanger Sequencing method (Michael and Jackson, 2013). In the early 2000s next-generation sequencing (NGS) technologies were being developed to improve the process of DNA sequencing. In 2005, 454-pyrosequencing, which uses light to detect nucleotides in a DNA sequence, was the first NGS platform to be released. When it was first released it could sequence 25 million base pairs with at least 99% accuracy in a single run. This was a 100-fold increase in output over the old Sanger method (Margulies et al., 2005). The Solexa Genome Analyzer was released in 2006. It used fluorescent dyes with reversible terminators to detect DNA sequence and was the first sequencer capable of sequencing a gigabase of data in a single run. The Illumina company acquired Solexa in 2007 and continued to develop the sequencing-by-synthesis method. Currently, the newest Illumina sequencers can produce 900-1800 Gigabases of data in single run and are capable of sequencing an entire human for nearly $1000, compared to the nearly $3 billion it took to sequence the first human genome. Ion Torrent is another platform for DNA sequencing that has been developed. It measures changes in pH when nucleotides are incorporated into a growing DNA molecule to accurately attain the sequence of bases in a DNA strand. All of the aforementioned sequencing technologies have one major disadvantage in that they all produce relatively short read lengths, 300-1000 bp. PacBio sequencing, another DNA

sequencing technology, uses single-molecule sequencing to create reads with an average length of ~10 kb with reads that are >40 kb in length. NGS has made genomic sequence data significantly cheaper and easier to produce. Indeed, as of July 2013, 55 plant genomes from 49 different species have been sequenced and assembled and are now available for use. Sequencing technologies continue to improve by becoming faster, less expensive, more reliable, and have higher output than ever before. As sequencing power increases developments in assemblers have also been necessary to keep up with the increase in genomic information. With this in mind, hundreds of sequenced and assembled plant genomes will become available in the next few years (Michael and Jackson, 2013). An increase in genetic information in plant species will unlock methods to develop higher yielding and more stress resistant crops to feed an ever growing human population.

Reference Genomes – An assembled genome is an invaluable resource for genetic study of any species of interest. It allows for expanded and more thorough studies that are not possible without access to a reference genome. Genome composition and repetitive elements, gene number and function, duplication events, domestication bottlenecks, and other genomic questions can be answered through the analysis of whole genome assemblies. Molecular breeding programs utilize reference genomes to identify genes and pathways involved in traits such as herbicide resistance, seed nutrition, and yield. These genes can then be selected for in order to provide more nutritious, higher yielding, and economically viable crops. The *Amaranthus* genus is lacking a high quality reference genome needed for the study of grain, vegetable, and weedy species of amaranths as scientists try to maximize the food qualities and minimize the invasive effects of both edible and weedy species. To date the only high quality

genome assembly in the Amaranthaceae family is the sugar beet (*Beta vulgaris*; Dohm et al., 2014).

Sugar Beet Genome Assembly – The sugar beet genome is the closest species to *A. hypochondriacus* with a sequenced genome and the first of the Caryophyllales, an order that consists of over 11,500 species. The sugar beet reference genome spanned 567 Mb (~75%) of the ~730 Mb estimated genome size. It was assembled into 2,171 scaffolds with 38,337 unscaffolded contigs and with 85% of the genome being assigned to one of the nine haploid chromosomes. Repeat content and description, gene annotation and prediction, and variation in accessions were also studied to further elucidate the genome composition and molecular pathways present in the sugar beet (Dohm et al., 2014).

Other Plant Reference Genomes – As mentioned previously, over 50 other plant genomes have been sequenced and assembled. The seven other genome assemblies papers (tomato, *Solanum lycopersicum*, peach, *Prunus persica*, cucumber, *Cucumis sativus*, cacao, *Theobroma cacao*, banana, *Musa acuminata*, Norway spruce, *Picea abies*, and apple, *Malus* x *domestica*) that were observed showed a remarkable amount of similarity to one other in their structure and genome studies performed (The Tomato Genome Consortium, 2012, The International Peach Genome Initiative, 2013, Huang et al., 2009, Argout et al., 2011, D'Hont et al., 2012, Nystedt et al., 2013, Velasco et al., 2010). In addition to a discussion of the sequencing and assembly of the genome, all of these papers also include discussion on the number of genes, specific genetic pathways or gene function, transposons and repeat content, and whole-genome duplications events (WGD) and how they relate to the evolutionary history of the species. Another common feature in these papers was exploring syntenic relationships between the newly assembled genome and published reference genomes of other plant species. This was done with the tomato, cucumber, and cacao

genomes. These common elements in genome assembly papers have made them formulaic and provide a framework for genome assembly papers in the future.

REFERENCES

Arendt, E. K., & Zannini, E. (2013). Wheat and other Triticum grains. *Cereal grains for the food and beverage industries*, (248), 1.

Alvarez-Jubete, L, Arendt, E.K., Gallagher E. (2009). Nutritive value and chemical composition of pseudocereals as gluten-free ingredients. *International Journal of Food Sciences and Nutrition*, *60*(4), 240-257.

Amicarelli, V., Camaggio, G. (2012). Amanranthus: A crop to rediscover. *Forum Ware International*, 2, 4-11.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408, 796-815.

Argout, X., Salse, J., Aury, J., Guiltiana, M., et al. (2011). The genomes of *Theobroma cacao*. *Nat Genet.*, 43(2), 101-108.

Baltensperger, D. D., Weber, L. E., & Nelson, L. A. (1992). Registration of 'Plainsman'grain amaranth. *Crop science*, *32*(6), 1510-1511.

Belitz, H.D., & Grosch, W. (1999). *Food Chemistry*, 2nd Ed., pp. 631–692, Springer-Verlag, Berlin, Heidelberg, Germany.

Bennett, M. D., & Smith, J. B. (1991). Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society B: Biological Sciences*,*334*(1271), 309-345.

Brenner, D.M. (1992). The Plainsman Story. *Legacy,* 5(1), 12-13.

Cai, Y.Z., Corke, H., Wu, H.X. (2004). Amaranth: *In:* Wrigley, C., Corke, H., and Walker, C (eds) *Encyclopedia of Grain Science*. Oxford: Elsevier

Caselato-Sousa, V.M., Amaya-Farfan, J. (2012). State of knowledge on amaranth grain: A comprehensive review. *Journal of Food Science*, 4, 93-104.

Chan, K.F., Sun, M. (1997). Genetic diversity and relationships detected by isozyme and RAPD analysis of crop and wild species of *Amaranthus*. *Theor Appl Genet* 95, 865–873.

Chauhan, B.S., Johnson, D.E. (2009). Germination ecology of spiny (*Amaranthus spinosus*) and slender amaranth (*A. viridis*): Troublesome weeds of direct-seeded rice. *Weed Science*, 57, 379-385.

Choi, H., Kim, W., Shin, M. (2004). Properties of Korean amaranth starch compared to waxy millet and waxy sorghum starches. *Starch*, 56, 469-477.

Culpepper, A.S., Grey, T., Vencill, W.K., Kichler, J.M., Webster, T.M., Brown, S.M., York, A.C., Davis, J.W., Hanna, W.W. (2006). Glyphosate-resistant Palmer amaranth (*Amaranthus pameri*) confirmed in Georgia. *Weed Science*, 54, 620-626.

D'Hont, A., Denoeud, F., Aury, J.M., Baurens, F.C., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, doi:10.1038/nature11241.

De la Rosa, A.P.B., Fomsgaard, I.S., Laursen, B., Mortensen, A.G., Olvera-Martinez, L., Silva-Sanchez, C., Mendoza-Herrera, A., Gonzolez-Castaneda, J., De Leon-Rodriguez, A. (2009). Amaranth (*Amaranthus hypochondriacus*) as an alternative crop for sustainable food production: Phenolic acids and flavonoids with potential impact on its nutraceutical quality. *Journal of Cereal Science*, 49, 117-121.

Délano-Frier, J. P., Avilés-Arnaut, H., Casarrubias-Castillo, K., Casique-Arroyo, G., Castrillón-Arbeláez, P. A., Herrera-Estrella, L., ... & Estrada-Hernández, M. G. (2011). Transcriptomic analysis of grain amaranth (Amaranthus hypochondriacus) using 454 pyrosequencing: comparison with A. tuberculatus, expression profiling in stems and in response to biotic and abiotic stress. *BMC genomics*, *12*(1), 363.

Dohm, J.C., Minoche, A.E., Holtgrawe, D., et al. (2014). The genome of the recently domesticated crop plant sugar beet (*Betavularis*). *Nature*, 505, 546-9.

Gamel, T.H., Linssen, J.P., Mesallam, A.S., Damir, A.A., Shekib, L.A. (2006). Effect of seed treatments on the chemical composition of two amaranth species: Pil, sugars, fibres, minerals and vitamins. *Journal of the Science of Food and Agriculture*, 86, 82-89.

Gamel, T.H., Mesallam, A.S., Damir, A.A., Shekib, L.A., Linssen, J.P. (2007). Characterization of amaranth seed oils. *Journal of Food Lipids*, 14, 323-334.

Gelinas, B., Seguin, P. (2007). Oxalate in grain amaranth. *Journal agricultural and food chemistry*, *55*(12), 4789-4794.

Greizerstein, E.J., Poggio, L. (1994), Karyological studies in grain amaranths. Cytologia 59, 25–30.

Gupta, V.K., Gudu, S. (1991), Interspecific hybrids and possible phylogenetic relations in grain amaranths. *Euphytica*, 52, 33-38.

He, H.P., Corke, H. (2003). Oil and squalene in *Amaranthus* grain and leaf. *J. Argic. Food Chem*, 51, 7913-7920.

Heap, I.M. (1997). The occurrence of herbicide-resistant weeds world-wide. *Pestic. Sci.,* 51:235-243.

Heap, I.M. (2004). International survey of herbicide resistant weeds. *Weed Science*, 68-70.

Huang, S., Li, R., Zhang, Z., Li, L., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. *Nature genetics*, *41*(12), 1275-1281.

The International Peach Genome Initiative. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet*. doi:10.1038/ng.2586.

Iturbide, G.A., Gispert, M. (1994), Grain amaranths (Amaranthus spp.). In: Hernandez-Bermejo JE, Leon J (eds) Neglected Crops: 1492 from a different perspective. FAO, Rome, pp 93–101.

Jimenez, F.R., Maughan, P.J., Alvarez, A., Kietlinski, K.D., Smith, S.M., Pratt, D.B., Elzinga, D.B., Jellen, E.N. (2013). Assessment of genetic diversity in Peruvian amaranth (*Amaranthus caudatus* and *A. hybridus*) germplasm using single nucleotide polymorphism markers. *Crop Science Society of America*, 53, 532-541.

Kauffman, C.S., Weber, L.E. (1990). Grain amaranth. In: J. Janick and JE Simon (eds.), Advances in new crops. Timber Press, Portland, OR p. 127-139.

Kietlinski, K.D., Jiminez, F., Jellen, E.N., Maughan, P.J., Smith, S.M., Pratt, D.B. (2014). Relationships between the weedy *Amaranthus hybridus* (Amaranthaceae) and the grain amaranths. *Crop Sci.*, 54, 220-228.

Knezevic, S.Z., Weise, S.F., Swanton, C.J. (1994). Interference of redroot pigweed (*Amaranthus retroflexus*) in corn (*Zea mays*). *Weed Science*, 568-573.

Konishi, Y., Nojima, H., Kazutoshi, O., Asaoka, M., Fuwa, H. (1985). Characterization of starch granules from waxy, nonwaxy, and hybrid seeds of amaranthus hypochondriacus L. *Agricultural and Biological Chemistry*, 49, 7, 1965-1971.

Macler, B.A., MacElroy, R.D. (1989). Productivity and food value of *Amaranthus creuntus* under non-lethal salt stress. *Adv. Space Res*,. *9*(8), 135-139.

Mallory, M.A., Hall, R.V., McNabb, A.R., Pratt, D.B., Jellen, E.N., Maughan, P.J. (2008). Development and characterization of microsatellite markers for the grain amaranths. *Crop. Sci.*, 48, 1098.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., ... & Volkmer, G. A. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376-380.

Martirosyan, D.M., Miroshnichenko, L.A., Kulakova, S.N., Pogojeva, A.V., Zoloedov, V.I. (2007). Amaranth oil application for coronary heart disease and hypertensions. *Lipids Health Dis.*, 6, 1.

Maughan, P.J., Sisneros, N., Meizhong, L., Kudrna, D., Ammiraju, J.S.S., Wing, R.A. (2008). Construction of an *Amaranthus hypochondriacus* bacterial artificial chromosome library and genomic sequencing of herbicide target genes. *Crop Science*, *48*(S1), 85-94.

Maughan, P.J., Yourstone, S., Jellen, E.N., Udall, J.A. (2009). SNP discovery via genomic reduction, barcoding and 454-pyrosequenceing in amaranth. *Plant Gen*., 2:260-270. doi:10.3835/plantgenome2009.08.0022.

Maughan, P.J., Smith, S.M., Fairbanks, D.J., Jellen, E.N. (2011). Development, characterization, and linkage mapping of single nucleotide polymorphisms in the grain amaranths (*Amaranthus* sp.). *The Plant Genome*, 4, 1-10.

Michael, T.P., Jackson, S. (2013). The first 50 plant genomes. *The Plant Genome*, *6*(2).

Moosavi, A., Afshari, R.T., Sharif-Zadeh, F., Aynehband, A. (2009). Effect of seed priming on germination characteristics, polyphenoloxidase, and peroxidase activities of four amaranth cultivars. *Journal of Food, Agriculture & Environment*, 7, 353-58.

Munns, R. (2002). Comparative physiology of salt and water stress. *Plant, Cell and Environment*, *25*(2), 239-50.

National Academy of Sciences. (1984). Amaranth: Modern prospects for an ancient crop. *Natl Acad Sci.*, Washington D.C.

Nystedt, B., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497, 579-584.

Oleszek, W., Junkuszew, M., Stochmal, A. (1999). Determination and toxicity of saponins from Amaranthus cruentus seeds. *J Agric Food Chem.*, 47, 3685–3687.

Omami, E.N., Hammes, P.S. (2006). Interactive effects of salinity and water stress on growth, leaf water relations, and gas exchange in amaranth (Amaranthus Spp.). *New Zealand Journal of Crop and Horticultural Science*, 34, 33-44.

Park, Y.J., Nemoto, K., Nishikawa, T., Matsushima, K., Minami, M., Kawase, M. (2011). Genetic diversity and expression analysis of granule bound starch synthase I gene in the new world grain amaranth (*Amaranthus cruentus* L.). *Journal of Cereal Science*, 53, 298-305.

Piskarikova, B., Kracmar, S., Herzig, I. (2005). Amino acid contents and biological value of protein in various amaranth species. *Czech J. Anim Sci.*, *50*(4), 169-174.

Qian, J.Y., Kuhn, M. (1999). Characterization of *Amaranthus creuntus* and *Chenopodium quinoa* starch. *Starch*, 51, 116-120.

Rastogi, A., Shukla, A.S. (2013). Amaranth: A new millennium crop of nutraceutical values. *Critical Reviews in Food Science and Nutrition*, 53, 109-125.

Reddy, L.H., Couvreur, P. (2009). Squalene: A natural triterpene for use in disease management and therapy. *Advanced Drug Delivery Reviews*, 61, 1412-1426.

Sauer, J.D. (1950). The grain amaranths and their relatives: a survey of their history and classification. *Ann Mo Bot Gdn,* 37, 561–619.

Sauer, J.D. (1967). The grain amaranths and their relatives: A revised taxonomic and geographic survey. *Ann Mo Bot Gdn*, 54, 103–137.

Sauer, J.D. (1993). Amaranthaceae—amaranth family. In: *Historical Geography of Crop Plants: A Select Roster*. CRC, Boca Raton, Florida, USA pp 9–14.

Saunders, R.M., Becker, R. (1984). Amaranthus: A potential food and feed resource. In Pomeranz Y (ed) *Advances in Cereal Science and Technology*, Vol 6. American Association of Cereal Chemists, St. Paul, Minn.

Stallknecht, G.F., Schulz-Schaeffer, J.R. (1993). Amaranth rediscovered. In: Janick, J., Simon, J.E. (eds), *New Crops*. Wiley, New York, pp 211–218.

Sunil, M., Hariharan, A.K., Nayak, S., Gupta, S., Nambisan, S.R., Gupta, R.P., Panda, B., Choudhary, B., Srinivasan, S. (2014). The draft genome and transcriptome of *Amaranthus hypochondriacus*: A C4 dicot producing high-lysine edible pseudo-cereal. *DNA Research*, 1-18.

Tapia, M. (1994). Neglected crops of the Andean region. In: Hernandez-Bermejo JE, Leon J (eds) *Neglected Crops: 1492 from a Different Perspective*. FAO, Rome, pp 144–148.

The Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485, 635-641.

Transue, D.K., Fairbanks, D.J., Robinson, L.R., Andersen, W.R. (1994). Species identification by RAPD analysis of grain amaranth genetic resources. *Crop Sci.*, 34, 1385–1389.

Tucker, J.B. (1986). Amaranth: The once and future crop. *Bioscience Biotechnology and Biochemistry*, 36, 9-13.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., et al. (2010). The genome of the domesticated apple (Malus [times] domestica Borkh.). *Nature Genetics*, *42*(10), 833-839.

Xu, F., & Sun, M. (2001). Comparative analysis of phylogenetic relationships of grain amaranths and their wild relatives (Amaranthus; Amaranthaceae) using internal transcribed spacer, amplified fragment length polymorphism, and double-primer fluorescent intersimple sequence repeat markers. *Molecular Phylogenetics and Evolution*, *21*(3), 372-387.

Yanez, G.A., Messinger, J.K.,Walker, C.E., Rupnow, J.H. (1986). Amaranthus hypochondriacus: Starch isolation and partial characterization. *Cereal Chem.*, 63, 273–276.