Expanded Understanding of *Eleusine* Diversity and Evolution

by

Hui Zhang

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 4, 2018

Key words: Next-generation sequencing, Illumina, Genome, Transcriptome, Phylogenetic, Goosegrass

Approved by

J. Scott McElroy, Chair, Professor of Crop, Soil and Environmental Sciences
Charles Y. Chen, Professor of Crop, Soil and Environmental Sciences
Leslie R. Goertzen, Associate Professor of Biological Sciences
Eric Peatman, Associate Professor of Fisheries, Aquaculture, and Aquatic Sciences

Abstract

Eleusine, including 9 to 12 species, is a small genus of annual and perennial grass species within the Eragrosteae tribe and Chloridoideae subfamily. There are very few genomic information about this genus. The primary goal of this dissertation research is to expand understanding of *Eleusine* diversity and evolution. The dissertation opens with a brief literature review regarding the motivation for this research. In chapter 2 we report a draft assembly of approximately 498 Mb whole-genome sequences of goosegrass obtained by *de novo* assembly of paired-end and mate-paired reads generated by Illumina sequencing of total genomic DNA. From around 88 GB of the clean data, the genome was assembled into 24,063 scaffolds with N50 = 233,823 bp. The nuclear genome assembly contains 25,467 predicted unique protein-coding genes. Sixteen target herbicide resistant genes and four non-target herbicide resistant gene families were obtained from this draft genome.

Chapter 3 presents a complete plastid genome sequence of goosegrass obtained by *de novo* assembly of paired-end and mate-paired reads obtained in chapter 2. The goosegrass plastome is a circular molecule of 135,151 bp in length, consisting of two single-copy regions separated by a pair of inverted repeats (IRs) of 20,919 bases. The large (LSC) and the small (SSC) single-copy regions span 80,667 bases and 12,646 bases, respectively. The plastome of goosegrass has 38.19% GC content and includes 108 unique genes, of which 76 are protein-coding, 28 are transfer RNA, and 4 are ribosomal RNA.

Finally, Chapter 4 presents a study utilizing transcriptome to resolve *E. coracana* heritage problem. We developed transcriptomes for six *Eleusine* species from fully developed seedlings using Illumina technology and three *de novo* assemblers (Trinity, Velvet, and SOAPdenovo2) with the redundancy-reducing EvidentialGene pipeline. *E. coracana* reads filtered for only nuclear-encoded genes were

mapped to *E. indica* transcriptome and the unmapped reads were extracted and assembled to create a *E. coracana* Synthetic B transcriptome. The other five *Eleusine* species' transcriptome reads were mapped to the *E. coracana* Synthetic B transcriptome, however, no mapped percentage exceeded 54.5%. By comparison, the reads of *E. indica* mapping to the *E. coracana* Synthetic A transcriptome is 72.9%. Variants and phylogenetic analyses found that no diploid *Eleusine* species close to the *E. coracana* Synthetic B transcriptome branch. The utilization of our synthetic B transcriptome is openly available as a resource to aid in the future identification of the paternal genome donor of *E. coracana*.

Acknowledgments

I would like to express my sincere appreciation to my great advisor, Prof. J. Scott McElroy, for his patience, inspiration and encouragement through my PhD studies. Without his guidance and support, none of my research in the past four years would be possible. Because of him, I have a very good project which I am very interested in and a very good committee. I am grateful to all of my dissertation committee members, Dr. Charles Y. Chen, Dr. Leslie R. Goertzen, and Dr. Eric Peatman, for their time, patience and advice related to my research work. I would like to thank Dr. Tonia S. Schwartz for agreeing to be my dissertation outside reader.

I would like to thank Dr. Shu Chen for his kindness help in both study and life when I just came to Auburn. Special thanks to Nathan D. Hall, for his expertise at Python and scripting, which made my work much more manageable. I would like to thank my friends Frank and Sandra for helping me all the time.

I am extremely grateful for China Scholarship Council, for their financial support, which made my study in the United States possible. I would like to thank our lab technician Dr. Jinesh Patel and all my lab mates including Mr. Bo Bi, Ms. Xiaoli Ma, Ms. Suma Basak, and Mr. Austin Brown.

Finally, but definitely not least, I would like to express my deepest thanks to my family, for all their love, encouragement and support.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 2

Chapter 3

Chapter 4

List of Abbreviations

ABC	ATP-binding Cassette
CYP450	Cytochrome P450
GST	Glutathione S-transferase
GT	Glycosyltransferase
SSR	Simple Sequence Repeat
LINE	Long Interspersed Nuclear Element
SINE	Short Interspersed Nuclear Element
LTR	Long Terminal Repeat-type
TSR	Target Site Resistance
NTSR	Non Target Site Resistance
NGS	Next Generation Sequencing
LSC	Large Single Copy
SSC	Small Single Copy
IR	Inverted Repeat
NPGS	National Plant Germplasm System
GRIN	Germplasm Resources Information Network
SNV	Single Nucleotide Variant
MNV	Multiple Nucleotide Variant
GO	Gene Ontology

Chapter 1. Introduction

Review of literatures

Genus Eleusine Gaertn. (Gramineae). *Eleusine* GAERTN, family Poaceae, is a small genus of annual and perennial grass species within the Eragrosteae tribe and Chloridoideae subfamily. It includes about 9 to 12 species that can hybridize to form intermediates and they are very similar in morphological feature (Hilu, 1981; Phillips, 1972; Mehra, 1962; Willis, 1973). It is mainly distributed in the tropical and subtropical parts of Africa, Asia and South America (Phillips, 1972). *Eleusine* contains diploid and tetraploid species, with chromosome numbers ranging from 2n = 16, 18 or 20 in diploids to 2n = 36 or 38 in tetraploids. All of the species are wild except *E. coracana*, which is cultivated for grain and fodder in Africa and the Indian subcontinent. The center of *Eleusine* diversity is East Africa and there are eight species in this genus occuring in this region, which includes *E. africana, E. coracana, E. kigeziensis, E. indica, E. floccifolia, E. intermedia, E. multiflora and E. jaegeri* (Mehra, 1963a; Phillips, 1972). The genome size of *Eleusine* species is very small and the 2C DNA amount ranges from 2.50 pg to 3.35 pg for diploid species (Hiremath and Salimath, 1991a). Questions remain regarding the evolutionary origins of the polyploid species and their relationship to weedy/wild diploid progenitors.

Goosegrass (Eleusine indica (L.) Gaertner). Goosegrass (*Eleusine. indica*) is an annual, diploid (2n = 2x = 18), self-pollinating grass species in *Poaceae* family (Zhang et al., 2017). A single plant can produce more than 50,000 small seeds easily dispersed by water and wind (Waterhouse, 1993). Goosegrass is primarily listed as an agricultural and environmental weed (Randall, 2012) and is considered a “serious weed” in at least 42 countries and also listed as one of the world’s five worst weed species due to its high reproductive capacity, herbicide resistance and wide tolerance to various environments (Holm et al., 1977; Chen et al., 2015). Once goosegrass is established it has a notoriously tough root system making it necessary to use physical or mechanical control. In addition, the resistance of

goosegrass to a wide range of herbicides has greatly compounded the difficulties with its control. According to the International Survey of Herbicide Resistant Weeds, a total of 25 resistance cases have been reported in goosegrass (Chen et al., 2015). In addition, there is also strong evidence that *E. indica* is the maternal genome donor of the domesticated crop species finger millet (*E. coracana*) (Bisht and Mukai, 2001a, b, Hilu, 1988 and Hiremath and Salimath, 1992). Besides, goosegrass as one of the worst agricultural weeds worldwide, it can decrease the crop yield and lead to huge economic loss (Ma et al., 2015); additionally, as a weed in turfgrass, it will most likely affect the health and beauty of turf and increase weed control problems (McElroy, 2016).

Finger millet (Eleusine coracana) and Eleusine. Africana. *E. coracana (L.).* GAERTN. subsp, *coracana*, which is cultivated as both grain and fodder in Africa and the Indian subcontinent. *E. coracana*, also named finger millet or African finger millet, is an allotetraploid species with chromosome number of 2n = 4x = 36, which is also an economically important minor crop that ranks third in cereal production in semiarid regions of the world (Bisht and Mukai, 2001b). *E. coracana* is drought and diseases resistant (Singh et al., 2014); easily digested and rich source of calcium, potassium, and fiber (Shobana et al., 2013); has high medicinal value (Bhatnagar, 1952); and is beneficial as animal fodder (Bisht and Mukai, 2002).

Finger millet is an orphan crop, an important regional crop that lacks widespread use and has minimal genetic and genomic resource (Singh et al., 2014). It is reported to have higher calcium and potassium content compared to traditional grains such as maize (*Zea mays*), rice (*Oryza sativa*), and wheat (*Triticum* spp.) (Singh and Raghuvanshi, 2012). Orphan crops are important sources of standing genetic variation within agriculture and if their use and development is not facilitated, we could lose vital genetic sources for crop improvement as well as the key parts of many unique and beautiful cultures. Orphan crops also have societal benefits of aiding to sustain cultural richness and maintain community identity in rural societies (Naylor et al., 2004). Greater consumption of industrially grown crops such as corn, wheat, and rice leads to less research attention given to orphan crops to improve varieties and

agronomic practices for improved production further diminishing the use (Hammer and Heller, 2001). Orphan crops such as finger millet could be a beneficial food sources to a ballooning world populations because they can be grown on more marginal land under harsher environmental conditions (Naylor et al., 2004).

Another *Eleusine* species, named *E. africana*, has been considered as a close relative to *E. coracana* and they have the same genome and chromosome number (2n=36) (Mehra 1962, Chennaveeraiah and Hiremath 1974a). It claimed that *E. africana* is a wild species, but *E. coracana* is a cultivated species. Most of the reports favor that two diploid species *E. indica* and *E. floccifolia* cross, then their chromosomes double, which produce the *E. africana*; then mutation occurs, which produces *E. coracana*. However, all of this is controversial and requires confirmation (Bisht and Mukai 2001a, b).

Weed genomics. Weed genomics is the study of the structure, function and evolution of weed genomes using genomic tools (Basu et al., 2004). Weed genomics can help us to understand weed biology and weediness, find new herbicide targets, elucidate targets of know herbicide, understand the mechanisms of evolved herbicide resistance, and find new weed control strategies (Peng et al., 2014). The genomics resources for weed research are still meager compared with those for crops and there is no good model for weed genomics until now (Stewart, 2009). There are relatively few weed genomes and transcriptomes available despite the rapidly developing technology. To our knowledge, only waterhemp (Lee et al., 2009) and horseweed (Peng et al., 2014) draft nuclear genomes have been published and no fully sequenced weed genome has been published. Goosegrass and crofton weed (*Ageratina adnophora*) chloroplast genomes have been assembled based on a high-throughput sequencing approach (Zhang et al., 2017; Nie et al., 2012).

Weeds evolve fast since it can survive very extreme conditions and few herbicides can control with time going on. Genomic method is a new way to increase our understanding of the evolution of herbicide resistance and the basic genetics that make weeds a successful group of plant. An important purpose of effective weed management is to stop or slow the evolution and spread of herbicide resistance in weeds.

There are mainly two genomic methods to study weed: genome analyses and traits analyses (Stewart et al., 2009). Genome analyses can provide loci and traits of interest and traits analyses can generate genes related with weedy traits. Besides, Non-target-site-resistance (NTSR) belongs to a quantitative trait and the complex genetic control of NTSR to three herbicides inhibiting acetylcoenzyme A carboxylase (ACCase) and one inhibiting acetolactate-synthase (ALS) was investigated in black-grass (Petit et al., 2010). Molecular markers and herbicide resistance loci can be achieved through NGS and used for weeds herbicide resistant study.

Next Generation Sequencing. Next-generation sequencing (NGS), known as high-throughput sequencing, is a non-Sanger-based high-throughput methodology that enables rapid sequencing of nucleotides in DNA or RNA samples in a very short time (Reis-Filho, 2009). It has been widely used in different areas since first introduced to the market in 2005 with its ultra-high throughput, scalability, and speed. NGS technologies are typically represented by Illumina (Solexa) sequencing, Roche 454 sequencing, Ion torrent PGM sequencing, SOLiD sequencing, and PacBio single-molecular sequencing (Liu et al., 2012). Each method has its own advantages and disadvantages. In recent years, NGS technologies with low cost and high throughput have dramatically accelerated the pace of generating new genomes and transcriptomes, which made it possible for groups and even individual investigators to sequence the genome of the species they study.

Genome assembly. Horseweed (*Conyza canadensis*) genome has been sequenced by multiple sequencing platforms (454 GS-FLX, Illumina HiSeq 2000, and PacBio RS) using various libraries with different insertion sizes (approximately 350 bp, 600 bp, 3 kb, and 10 kb) and assembled by SOAPdenovo and CLC Genomic Workbench (Peng et al., 2014). Lee et al. (2009) sequenced and assembled the waterhemp (*Amaranthus tuberculatus*) genome using 454-pyrosequencing and GS Assembler (Newbler) (Lee et al., 2009). Draft *Eleusine coracana* nuclear genome has been reported by Hittalmani et al. (2017) using Illumina and SOiLD sequencing technologies (Hittalmani et al., 2017). Hatakeyama et al. (2017)

used diverse technologies with sufficient coverage and assembled it via a novel multiple hybrid assembly workflow that combines next generation with single-molecule sequencing (Hatakeyama et al., 2017). *Oropetium thomaeum* genome sequenced by PacBio platform (VanBuren et al., 2015), and sweet orange genome sequenced by Illumina GAII sequencer (Xu et al., 2013). Although more and more plant genomes have been sequenced and published, there is no fully finished genome. However, it is still challenge to produce a quality genome because of gene duplications, repeat regions, non-uniform coverage of the target and sequencing bias (Pop and Salzberg, 2008). Besides, plants genome usually have large genome size, higher ploidy, high rates of heterozygosity and repeats, and complex gene contents (Schatz et al., 2012).With the price and error rate become lower of PacBio sequencing, it is possible to sequence plant genomes with higher accuracy and coverage.

Transcriptome assembly. RNA-Seq is becoming more and more popular with the price decreasing compared to microarrays. It has been proved useful in different various areas, especially in non-model species where genetic and genomic resources are limited or unavailable (Góngora-Castillo and Buell, 2013). Transcriptome sequencing can be finished in a very short time, however, assembling it without a known reference remains challenging since the sequencing reads are usually very short. Until now, several software used for *de novo* transcriptome assemble have been developed, such as Trinity (Grabherr et al., 2011; http://trinityrnaseq.github.io), Velvet (Zerbino and Birney, 2008; https://www.ebi.ac.uk/~zerbino/ velvet/), Trans-ABySS (Robertson et al., 2010), and SOAPdenovo (Luo et al., 2012; https://github.com/aquaskyline/SOAPdenovo2). Besides, N50 and BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015) can be used to check the quality and completeness of assembly.

An et al. (2014) sequenced the transcriptomes of two goosegrass biotypes using Illumina Genome platform and assembly analysis using Trinity *de novo* assembler produced 158,461 transcripts and 100,742 unigenes with an average length of 712.79 bp (An et al., 2014). Chen et al. (2015a) used three different *de novo* assemblers (Trinity, Velvet, and CLC) and an EvidentialGene pipeline tr2aacds to

assemble two optimized transcript sets for goosegrass (Chen et al., 2015a). *E. coracana* transcriptome has been sequenced and assembled using different methods (Kumar et al., 2014; Rahman et al., 2014; Hittalmani et al., 2017).

Plastid genome assembly. Plastid genome sequencing is an essential tool to study plant evolution. The small size and highly conserved of chloroplast genome makes it suitable and invaluable for complete sequencing and phylogenetic analysis (Cho et al., 2015). The first complete plastid genome sequences were published in 1986 (Maier and Schmitz-Linneweber, 2004). Traditional methods to sequence the chloroplast genome or partial chloroplast genes rely on costly and time-consuming plastid isolation, PCR and amplicon sequencing. With the advent of next-generation sequencing (NGS) technology, new approaches for chloroplast genome sequencing have been gradually proposed due to their high-throughput, time-saving and low-cost (Cronn et al., 2008). The number of available complete chloroplast genomes has increased rapidly due to high-throughput sequencing technology. However, it is still challenging to assemble plastid genome since some genes might have integrated into the nuclear DNA through horizontal gene transfer. Zhang et al. (2017) reported the complete plastid genome sequence of goosegrass obtained by *de novo* assembly of paired-end and mate-paired reads generated by Illumina sequencing of total genomic DNA. The goosegrass plastome is a circular molecule of 135,151 bp in length, consisting of two single-copy regions separated by a pair of inverted repeats (IRs) of 20,919 bases (Zhang et al., 2017).

Phylogenetic analyses. Phylogenetic tree is a popular method to present the phylogenetic relationships among species. However, the first step usually need to alignment nucleotides or the corresponding amino acid sequences of homologous. Low copy nuclear genes, chloroplast genes, and mitochondrial genes can be used in phylogenetic analyses since they evolved slowly. The efficiency of the transcriptome by Illumina sequencing can ensure a thorough analysis of chloroplast genes, mitochondria genes and conserved nucleus genes. Chloroplast genes are good for phylogenetic studies since they are highly

conserved compared with other genes (Hilu, 1988) and mitochondria genes used in phylogenetic studies also have been documented in several studies (Meyer and Wilson, 1990; Zardoya and Meyer, 1996).

Hatakeyama et al. (2017) constructed a molecular phylogenetic analysis using the detected low-copy-number homeologs during *E. coracana* genome assembly, and *E. indica* was close to *E. coracana*, which is consistent with our results (Zhang et al., 2018, unpublished). *Eleusine* species phylogenetic relationships and maternal ancestry have been solved by amplifying seven chloroplast genes/intergenic spacers (*trnK, psbD, psaA, trnH–trnK, trnL–trnF, 16S and trnS–psbC*) (Agrawal et al., 2013). Phylogenetic tree was produced using a 1.1-kb region of the cytochrome c oxidase subunit I (*coxI*) gene to investigate possible changes in diatom mitochondrial genetic codes (Ehara et al., 2000). Phylogenetic relationships in the genus *Eleusine* (Poaceae: Chloridoideae) were also investigated using nuclear *ITS* and plastid *trnT–trnF* sequences (Neves et al., 2004).

Herbicide resistance. There are currently 494 unique cases of herbicide resistant weeds globally until 2018, which include 148 dicots and 106 monocots (Heap, 2018). Herbicide resistant weeds have been reported in 92 crops in 70 countries and weed have evolved resistance to 163 different herbicides with 23 of 26 know herbicide sites of action (Heap, 2018). Weeds evolve fast since they can tolerate very extreme conditions and few herbicides can control with time going on. Resistant weeds can survive herbicide application by a variety of mechanisms, which are now mainly including two categories: target-site resistance (TSR) and non-target-site resistance (NTSR) (Yuan et al., 2007). NTSR is still not well defined at the genomic or molecular level for most weeds, although it is interesting both biologically and practically (Peng et al., 2014). Development of genomic resources and approaches can expedite research of these two herbicide resistance mechanisms for most of weeds.

Goosegrass has many different populations and some are resistant to multiple herbicide modes of action. Numerous goosegrass populations throughout the world resistant to: glyphosate (EPSP synthase inhibitors); pendimethalin, prodiamine, trifluralin (microtubule inhibitors), glufosinate (glutamine synthase inhibitors); fluazifop, fenoxaprop, haloxyfop, sethoxydim, clethodim (ACCase inhibitors);

paraquat (PSI inhibitors); metribuzin (PSII inhibitors); and imazapyr (ALS inhibitors) (McElroy, 2016; Zhang et al., 2017).

Dissertation objective

This dissertation mainly seeks to accomplish four goals: 1) Draft goosegrass genome assembly and application; 2) Goosegrass chloroplast genome assembly; 3) Construct high-quality transcriptome references for six *Eleusine* species and *E. coracana* synthetic B genome donor; 4) Use transcriptome sequencing to determine *E. coracana* heritage.

Chapter 2. Draft goosegrass (*Eleusine indica*) genome assembly and application

Introduction

Next generation sequencing technologies with low cost and high throughput have dramatically accelerated the pace of generating new genomes and transcriptomes, which made it possible for groups and even individual investigators to sequence the genome of the species they study. To date, transcriptomes using next-generation sequencing (NGS) of some weed species have been produced to find the mutation site of genes that contribute to herbicide resistance (McCullough et al., 2016a; b; Tehranchian et al., 2016). Transcriptome (Chen et al., 2015), chloroplast genome (Zhang et al., 2017) and mitochondria genome (Nathan et al., 2018; under review) assemblies have been produced for goosegrass and 78 plastid protein coding loci were sequenced for *E. coracana* (Givnish et al., 2010). Recently, genome and transcriptome assemblies also have been reported for *E. coracana* (Hittalmani et al., 2017; Hatakeyama et al., 2017) and some other *Eleusine* species' transcriptomes including *E. coracana* were also assembled (Zhang et al., 2018; unpublished).

Weed caused $36 billion annual damage in the United States alone, which included $32 billion in lost of crop production and $4 billion on herbicides that applied to crops (Pimentel et al., 2012). Resistant weeds can survive herbicide application by a variety of mechanisms, which can be classified into two categories: target-site resistance (TSR) and non-target-site resistance (NTSR) (Yuan et al., 2007). NTSR is still not well defined at the genomic or molecular level for most weeds (Peng et al., 2014). Development of genomic resources and approaches can expedite research of these two herbicide resistance mechanisms for most of weeds. All of the genes including herbicide resistant genes can be found from genome or transcriptome, which can then be used to develop the new method or new herbicide to control weeds. Herbicide resistance is absolutely the most important trait affecting long-term

control of weed populations. Genomics method can help to understand weed environmental adaptation, the evolution of herbicide resistance, elucidate the action of herbicides and dissect the herbicide resistant genes (Steward et al., 2009). Thus, the pathways of herbicide response may also suggest new molecular targets for herbicide development and novel weed management strategies. However, until now only two draft weed genomes have been sequenced: waterhemp (Lee et al., 2009) and horseweed (Peng et al., 2014). Unlike crops, the ultimately goal for weed scientist is to manage and control weed and reduce economic cost, so a good draft genome is enough right now for weed science. Weedy genes and significant markers can be found from a weed draft genome through genomes comparation between weeds and crops. Those weedy genes might be used to study herbicide resistance or develop new herbicides. New target of herbicide or receptors related to action site for herbicide may identified from genome sequencing. In addition, the conserved gene sequences in one weedy species can be used to amplify and clone by other weedy species since many genes are conserved. Such as the target site of action for ALS inhibiting herbicides was first cloned from Arabidopsis (Haughn et al., 1988). Then, these sequences were subsequently used to design primers for amplification and cloning homologous genes from many other weedy species (Horvath, 2009). Using draft genome databases should allow cloning of numerous genes of interest by weed biologists. Besides, it is necessary to know gene sequence to amplify the gene and check its expression.

Goosegrass as one of the worst agricultural weeds worldwide, it can decreases the crop yield, such as it can result in a 50% cotton yield loss (Ma et al., 2015). Additionally, as a weed in turfgrass, it decreases turf density and aesthetic value and increases weed control problems (McElroy, 2016). Relative to other weed species little is known regarding the biology of goosegrass, especially the biology of populations that have adapted to turfgrass management conditions. To date, goosegrass has evolved resistance to several herbicides, including the glyphosate (EPSP synthase inhibitors); pendimethalin, prodiamine, trifluralin (microtubule inhibitors), glufosinate (glutamine synthase inhibitors); fluazifop, fenoxaprop, haloxyfop, sethoxydim, clethodim (ACCase inhibitors); Paraquat (PSI inhibitors); metribuzin (PSII inhibitors); and imazapyr (ALS inhibitors) (McElroy, 2016; Heap, 2018). Besides, goosegrass has been

proved as maternal genome donor for *E. coracana*, which is an important allotetraploid minor crop species cultivated as both grain and fodder in Africa and the Indian subcontinent. Thus, the construction of a comprehensive draft genome should not only facilitate goosegrass herbicide resistance, but also benefit to study other *Eleusine* species, especially *E. coracana*.

Most of the cases of resistance of goosegrass to herbicides are due to change of the site of action (Heap, 2016). Goosegrass is a unique species that is diploid, is phenotypically diverse and adaptable to turfgrass and crop management practices, and has rapid evolution to herbicide resistance, which makes it easier to study than polyploidy. In this study, using Illumina genome sequencing technology, we have generated and assembled a draft genome sequence for goosegrass with an assembled N50 contig size reaching 210, 958 bp and N50 scaffold size of 233,823 bp. To our knowledge, this is the first draft genome of goosegrass. We also carried out several analyses using the complete sequence data, including genome contents, repeat elements, synteny analyses, and genome evolution and application. The primary goal of this project was to construct substantial genomic resources for goosegrass, which will be useful to understand the genomic basis of weediness traits, herbicide resistance and the evolutionary biology of this highly adaptable weed species.

Materials and methods

Plant materials and DNA extraction. The goosegrass population used in this research has been previously utilized for transcriptomic research (Chen et al., 2015). We refer to this goosegrass biotype as PBU, as it was collected from the E.V. Smith Research and Education Center-Plant Breeding Unit (PBU) of Auburn University. We consider this biotype as a holotype because it possesses typical characteristics of goosegrass grown in row-crop agricultural settings. Previously collected seeds were grown in potting medium (Miracle-Gro Potting Soil, Scotts Miracle-Gro Products, and Marysville, OH) to allow uniform germination. Four weeks later, seedlings were transplanted to plastic pots (10 cm × 10 cm × 8.5 cm) containing a native Wickham sandy loam soil with pH 6.3 and 0.5% organic matter. All plants were

seeded and grown in Auburn, AL (32.35°N, 85.29°W) in a glasshouse at 23 ± 2°C, and 71% average relative humidity. Total genomic DNA was extracted from fresh leaves using the DNeasy Plant Mini Kit (Qiagen, CA, USA).

Next generation sequencing. 1) DNA library preparation. At GENEWIZ, DNA samples were quantified using Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and the DNA integrity was checked with 0.6% agarose gel with 50-60 ng sample loaded in each well. DNA library preparations and sequencing reactions were conducted at GENEWIZ, Inc. (South Plainfield, NJ, USA). NEB NextUltra DNA Library Preparation kit was used following the manufacturer's recommendations (Illumina, San Diego, CA, USA). Briefly, the genomic DNA was fragmented by acoustic shearing with a Covaris S220 instrument. The DNA was then end repaired and adenylated. Adapters were ligated after adenylation of the 3'ends. Adapter-ligated DNA were indexed and enriched by limited cycle PCR. DNA libraries were validated using a DNA Chip on the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and were quantified using Qubit 2.0 Fluorometer. At the Auburn University Genomic and Sequencing Lab (http://www.ag.auburn.edu/enpl/gsl/seq.php), DNA libraries were constructed using Illumina's TruSeq Stranded DNA Sample Preparation Kit (Illumina, San Diego, CA USA) and library quantification was performed using the Kapa quantification kit for next generation sequencing with the Illumina platform (Kapa Biosystems, Wilmington, MA USA).

2) Mate pair library preparation. At GENEWIZ, the Illumina Nextera Mate Pair Library Prep kit (Illumina, San Diego, CA, USA) was used for the generation of the mate pair library. Briefly, 4 ug of high molecular weight DNA was fragmented and adapter-tagged by transposon. Biotin label nucleotides were introduced into DNA fragments by strand displacement reaction. DNA fragments were then separated on agarose gel and 5-7 kb fragments were recovered. Recovered DNA fragments were circularized and sheared with Covaris machine. Biotinylated "mate-pair" fragments were enriched by streptavidin bead, followed by end-repair, A-tail and adapter ligation to generate the sequencing library. DNA libraries were

validated using a DNA Chip on the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and were quantified using Qubit 2.0 Fluorometer.

3) HiSeq sequencing. At GENEWIZ, the DNA libraries were quantified by real time PCR (Applied Biosystems, Carlsbad, CA, USA), and multiplexed in equal molar mass. The pooled DNA libraries were clustered onto a lane of a flowcell, using the cBOT from Illumina. After clustering, the samples were loaded on the Illumina HiSeq 2500 instrument according to manufacturer's instructions. The samples were sequenced using a 2x 150 paired-end (PE) Rapid Run configuration. Image analysis and base calling were conducted by the HiSeq Control Software (HCS) on the HiSeq2500 instrument. Additional sequencing was conducted at Auburn University Genomic and Sequencing Lab, samples were loaded per lane and randomly distributed within and across lanes. The flowcells were placed in the HiSeq 1500 sequencer (Illumina Inc, San Diego, CA USA) and fluorescently labeled bases were attached to the complementary bases of each sequence. Next generation sequencing was performed with 100 bp PE read single indexing index sequencing protocol in Rapid Mode. Sequencing data was de-multiplexed with CASAVA software 1.8.2 (Illumina, San Diego, CA USA).

Sequence data analysis, assembly, and annotation. Raw reads quality were checked by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then processed by Trimmomatic v.0.33 (Bolger et al., 2014; http://www.usadellab.org/cms/?page=trimmomatic) to remove adaptors and low qualified reads and sequences. The trimmed reads were evaluated with FastQC again and normalized with Trinity's in silico read normalization (http://trinityrnaseq.github.io), with maximum coverage of 30. *De novo* assembler Velvet 1.2.08_ maxkmer101 (Zerbino and Birney, 2008; https://www.ebi.ac.uk/~zerbino/velvet/) were used with k-mer size 61 to get contigs. The contigs were further scaffolded using SSPACE standard v3.0 (Boetzer et al., 2011), which is a stand-alone scaffolder of pre-assembled contigs using paired-read data and gaps in the scaffolds were closed by GapCloser v1.12 (Luo et al., 2012).

The quality and genome completeness of assembly was checked by BUSCO (Benchmarking Universal Single-Copy Orthologs) v3 (Simão et al., 2015). N50s and contig length distributions of the assemblies were calculated with the script Count_fasta.pl (http: //wiki.bioinformatics.ucdavis. edu/index.php/Count_fasta.pl). Reads mapping was conducted using the tools 'map reads to reference' and 'probabilistic variant detection' separately in CLC Genomics Workbench 6.5.2 (CLC Bio, Aarhus, Denmark, http://www.clcbio.com) and BWA v0.7.5a (Li, 2013). The mapping parameters were set to 'Mismatch cost=3, Insertion cost=3, Deletion cost=3, Length fraction=0.95, Similarity fraction=0.95'.

RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) was used to build a new repeat library based on the goosegrass genome. The new library was blast with Uniprot and NCBI non-redundant (Nr) database to filter protein-coding sequences. Then, the filtered results were used to construct a new library for RepeatMasker and RepeatMasker v3.2.7 (http://www.repeatmasker.org/) was used to find homolog repeats in the genome. Identified repeats were classified into different known classes as per standard genome analysis.

AUGUSTUS (http://www.softberry.com) was used for genome annotation using *Zea mays* as a reference gene model. Besides, RNA-seq data was incorporated to AUGUSTUS to assist the gene prediction. Then, the amino-acid sequences predicted from AUGUSTUS were used as queries against the Uniprot and NCBI non-redundant (Nr) database with ncbi-blast-2.2.28+ with an E-value threshold of le-5. The blast results were processed to Blast2GO to retrieve GO terms. KEGG analyses also did in Blast2GO.

The protein sequences of rice (http://floresta.eead.csic.es/rsat/data/genomes/Oryza_sativa.IRGSP-1.0.29/genome/Oryza_sativa.IRGSP-1.0.29.pep.all.fa), foxtail millet (ftp://penguin.genomics.cn/pub/10.5524/100001_101000/100020/Millet.fa.glean.pep.v3.gz), brachypodium (ftp://ftp.ensemblgenomes.org/pub/plants/release31/fasta/brachypodium_distachyon/pep/), maize (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/pep/), and sorghum (ftp://ftp.ensemblgenomes.org/pub/plants/release-31/fasta/sorghum_bicolor/pep/) were downloaded. The orthologs and paralogs among rice, foxtail millet, brachypodium, maize and sorghum were identified

using Orthofinder v1.1.8 (Emms and Kelly, 2015). Phylogenetic relationship of seven Poaceae species was obtained based on single copy ortholog genes. Trees were visualized with Figtree (Rambaut, 2009). The synteny relationship between the longest 50 scaffolds of goosegrass genome and other five Poaceae species was obtained by Symap v4.2 (Soderlund et al., 2011).

Analysis of herbicide resistance genes and genetic markers. Herbicide resistant genes including target and non-target families (ABC, CYP450, GS, and Glycosyl-transferases) were downloaded from NCBI database. Goosegrass herbicide resistant genes were retrieved through blast with its draft genome.

Repeat sequences were identified using microsatellite identification tool (Thiel, 2003), and each repeat sequence was ≥10 bp. Repeat sequences whose repeating sequence units were 2 bp with at least six times and arranged from 3 to 6 bp and repeated not less than five times were considered as SSRs. Repeat sequences with lengths ≥14 bp were considered as long repeat sequences. Two SSRs with interruption lengths each ≤100 bp were considered as compound microsatellite repeats.

Results

Genome sequencing and assembly. A total of 585,760,134 raw paired reads (100 bp length) were obtained from sequencing (Figure 2) with average coverage around 76 and 507,713,623 reads were kept after quality control. The assembly process resulted 34,415 contigs and the total length was 477,398,776 bp. The largest contig size was 2,426,064 bp, and the shortest contig size was 200 bp. The average contig size was 13,872 bp and the N50 value of contigs was 210,958 bp. After scaffolding and gaps filling, the totals of 497,777,567 bp scaffolds were generated and the N50 was 233,823 bp. The longest scaffold is 2,439,579 bp and the average scaffold length was 20,458 bp (Table 1). The genome size was estimated to be around 584 Mb by cytogenetic methods (Hittalmani et al., 2017). Compared with the estimated genome size, the final scaffolds of 497 Mb represented 86% of the total goosegrass genome. The goosegrass genome possesses a GC content of 40.83%, which is slightly higher than other sequenced

weed genomes (horseweed, 34.9%) (Peng et al., 2014) but slightly lower than *Eleusine. coracana* (44.76%) (Hittalmani et al., 2017).

To confirm the correctness of the final scaffold datasets, all of the sequencing reads were mapped to the assembled draft genome. More than 90% of the Illumina reads were mapped to the final scaffolds. BUSCO showed that these assemble had > 95% (1,368 complete single-copy orthologs of 1,440 total BUSCO groups) of complete single copy orthologs (Figure 3). These results suggested that our final scaffolds cover most of the *E. indica* genome. The raw sequencing reads are available in the NCBI Short Read Archive (SRA) at http://www.ncbi.nlm.nih.gov/ as accession SAMN09001275. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession QEPD00000000. The version described here is version QEPD01000000.

Genome annotation. Total of 25,467 genes were predicted in goosegrass genome based on *de novo* method of gene prediction using AUGUSTUS. The total of 3,180 GO terms were assigned to goosegrass genome, which include 429 GO terms with cellular component in 8,344 sequences, 1,568 GO terms with biological process in 11,954 sequences, and 1,183 GO terms with molecular function in 14,837 sequences (Figure 4). The relationships of GO terms among each component were also showed (Figure 5). For cellular component, the number of GO terms related with integral component of membrane was the largest in the genome, followed by nucleus and membrane. Oxidation-reduction process was the most in biological process component compared with others, followed by protein phosphorylation and regulation of transcription, DNA-templated. For molecular function component, ATP binding is the most one when compared with other GO terms (Figure 4). There were 129 metabolic pathways obtained during KEGG analyses and some of these pathways were related with herbicide resistant genes, such as valine, leucine and isoleucine biosynthesis pathway. 456 sequences were related with Purine metabolism pathway, which referred to 35 enzymes and this pathway included the highest number of sequences, followed by thiamine metabolism and biosynthesis of antibiotics pathways. In addition, there were 6 classes of main enzymes:

hydrolases (1087 sequences), transferases (909 sequences), oxidoreductases (463 sequences), lyases (142 sequences), isomerases (111 sequences), and ligases (89 sequences) (Table 2).

Repeat content in the draft genome. Repetitive DNA, which can affect genetic diversity, gene duplication and genome stability, is an important part in plant genome (Mahesh et al., 2016). A total of 107,512,057 bp (21.8% of the draft genome) was identified in goosegrass genome. Retrotransposon elements were the largest mobile elements found in the genome (15.3%), of which most were long terminal repeat-type (LTR) retroelements (13.7%), followed by LINEs (1.4%) and SINEs (0.2%). Except for retrotransposon elements, the total interspersed repeats also include DNA transposon elements (2.9%) and unclassified repeats (2.9%). However, the tandem repeats only include 14 satellites, 4,033 small RNA structures, 25,545 simple repeats, and 46,289 low complexity repeats (Table 3).

Identification of microsatellites in goosegrass genome. To get more genomic resources, SSRs in goosegrass genome sequences were identified and analyzed. In total, 115,417 simple SSRs were obtained using MISA. 3,881 sequences contain more than one SSR and 10,076 SSRs present in compound formation (Table 4). Most of the simple SSRs was monomer repeats occupying 65.1% of total SSRs, followed by 20.5% dimers (23,640), 13.3% trimers (15,307), 0.8% tetramers (930), 0.3% pentamers (284), and 0.1% hexamers (161) repeats (Figure 6A). Among monomer repeats, 78.0% were poly A/T (58577 microsatellites), and 22.0% were G/C (16518 microsatellites) types. Among dimers, AG/CT (61.1%) were highest followed by AT/AT (19.0%), AC/GT (16.9%), and CG/CG (3.0%). Similarly, CCG/CGG (17.2%) were highest among trimers followed by AGG/CCT (15.6%), AAG/CTT (14.6%), ATC/ATG (14.3%), AAC/GTT (9.3%), AGC/CTG (8.3%), AAT/ATT (8.2%), ACC/GGT (7.0%), ACG/CGT (3.6%), and ACT/AGT (2.0%) (Figure 6B). In case of tetramers type, AAAT/ATTT (20.7%) and AAAG/CTTT (18.9%) were higher as compared to other types of tetramers type repeats. The maximum number of pentamers and hexamers were of AAAAG/CTTTT type and AACACC/GGTGTT type, separately. In total, the monomer, dimer and trimer SSR types were high as compared to tetramer,

pentamer and hexamer SSR types in goosegrass genome. The similar observation was made in another *Eleusine* species *E. coracana* (Hittalmani et al., 2017). Therefore, SSRs identified in this study will have an important value in population genetics and evolutionary analysis.

Except for SSRs, repeat sequences with lengths ⩾14 bp were considered as long repeat sequences in goosegrass genome, and 116,940 long repeat sequences were detected (Table 4). The number of long repeat sequences great than 100 were summarized in Figure 7.

Comparative genomics and the evolution of goosegrass genome. Comparative genomics can provide a method to unravel the relationship between genomes by describing conserved chromosomes or chromosome regions between related species. 50 longest scaffolds (3911 unique genes) from goosegrass draft genome were compared with fully sequenced five Poaceae species: brachypodium, foxtail millet, maize, rice and sorghum. In total, we identified 136 synteny blocks between goosegrass and brachypodium; 155 between goosegrass and foxtail millet; 179 between goosegrass and maize; 156 between goosegrass and rice; and 148 between goosegrass and sorghum (Table 5). Among those five species, maize and goosegrass have more conserved genomic regions than other species (Figure 8). However, only the longest 50 scaffolds were used because there is no goosegrass linkage map. Previous study revealed more number of conserved genomic regions between *Eleusine coracana* and those five Poaceae species' genomes (Hittalmani et al., 2017). This synteny relationship of goosegrass with other plants will enable us to construct goosegrass linkage map in future and identify novel genes. In addition, phylogenetic analysis showed goosegrass closer with *Eleusine coracana*, followed by rice and brachypodium (Figure 9). In addition, 36,160 orthogroups and 243 single-copy orthogroups were found by Orthofinder in seven Poaceae genomes and there were 696 species-specific orthogroups (Table 6). 10,932 orthogroups include all of the seven species and 31.4% genes in orthogroups. Most of orthogroups only include two species, followed by seven species (Figure 9). The single-copy orthologue genes will be useful for phylogenetic study.

Find and analysis herbicide resistant genes. In the goosegrass genome, sixteen target herbicide resistant genes were extracted through BLAST (Table 7). Among those genes, ALS, psbA and two copies of vlcfa have no introns. ALS, HPPD, dihydropteroate synthase gene, EPSPS, fatty acid synthase gene, and psbA only have one copy in goosegrass genome. In addition, ACCase was prove to have more exons than other genes, which number is 28. Those information will be helpful for people to study herbicide resistant genes' structure and function and mutation mechanisms.

Gene families that are commonly associated with non-target herbicide resistance include cytochrome P450s, glutathione S-transferases (GSTs), glycosyltransferases (GTs), and ATP-binding cassette (ABC) transporters (Yuan et al., 2007). Several family members can hit to one same gene when blast because some family members are very similar. There were 61 unique ABC transporter genes corresponding to 158 family members. Also, 85 cytochrome P450s genes with 310 family members were found in goosegrass. In addition, there were 152 GT genes and 20 GS genes, which corresponds to family members with 370, 52, separately (Table 8).

Discussion

In this work, we sequenced and assembled a draft reference goosegrass genome sequence using Illumina paired-end and mate-paired sequencing method. This is the first reference genome sequence from goosegrass, a maternal genome donor of African finger millet, and will be valuable for weeds evolution and herbicide resistance study. With the advent of next generation sequencing, more and more plant genomes have been sequenced and assembled. For example, horseweed genome was sequenced by 454 GS-FLX, Illumina HiSeq 2000, and PacBio RS sequencing method (Peng et al., 2014), *Oropetium thomaeum* genome by PacBio platform (VanBuren et al., 2015), and sweet orange genome by Illumina GAII sequencer (Xu et al., 2013). However, it is still challenge to produce a complete genome because of gene duplications, repeat regions, non-uniform coverage of the target and sequencing bias (Pop and

Salzberg, 2008). Plants genomes usually have large genome size, higher ploidy, high rates of heterozygosity and repeats, and complex gene contents (Schatz et al., 2012). Mixed library with different insert length and high-coverage are two main factors to affect produce a fully sequenced genome. With the price and error rate become lower of PacBio sequencing, it is possible to sequence plant genomes with higher accuracy and coverage. In this study, we only used Illumina HiSeq and got a quality draft genome. Our methods to generate and assemble a draft sequence for an entire plant genome can be used to generate many more plant draft genome sequences with a fast and cost-effective manner. Goosegrass chloroplast and mitochondrial genomes have also been assembled separately using our sequenced data (Zhang et al., 2017; Hall et al., 2018, under review).

Goosegrass has a relatively small genome around 600 Mb (Hittalmani et al., 2017). To our knowledge, this is the first genome for goosegrass, second genome from *Eleusine*, and third genome for weeds. Until today, only waterhemp (Lee et al., 2009) and horseweed (Peng et al., 2014) draft genomes have been published and no fully sequenced weed genome has been published. Our comparative analysis showed that goosegrass had higher genomic synteny with maize, but we only used the longest 50 scaffolds for goosegrass. Besides, the transposable-element activity in the repetitive regions also can have effects on synteny block results. Reshuffling of short DNA segments by mobile elements can remove large-scale collinearity in heterochromatic regions. Collinearity and synteny can be used to check the conservation between species and thus to help to take better advantage of new genomic resources (Tang et al., 2008). The shared synteny blocks in this research can be used to construct goosegrass linkage map, predict the position of genes conferring key weediness traits in future, and study the chromosome reshuffling (Salse et al., 2004). Target and non-target herbicide resistant genes and gene families have been achieved from goosegrass, which will be useful to study goosegrass herbicide resistance and population evolution in future. As well as providing a reference genome for use in future research, we were also able to provide 115,417 simple SSRs, which will be important for population genetics (He et al., 2012) and phylogenetic analysis (Melotto-Passarin et al., 2011; Nie et al., 2012).

There are currently 490 unique cases of herbicide resistant weeds globally until 2018, which include 148 dicots and 106 monocots. Herbicide resistant weeds have been reported in 92 crops in 70 countries and weed have evolved resistance to 163 different herbicides with 23 of 26 know herbicide sites of action (http://www.weedscience.org/). Weeds evolve fast since it can tolerate very extreme conditions and few herbicides can control with time going on. Genomic method is a new way to increase our understanding of the evolution of herbicide resistance and the basic genetics that make weeds a successful group of plant. Target-site-based resistance (TSR) seems easy to study since it was controlled by a single gene (Jasieniuk et al., 1996; Darmency, 1994). However, non-target-site-based-resistance (NTSR) was complex to study due to the mechanisms are unpredictable. There were many family members for each NTSR family in goosegrass genome. Fortunately, NTSR belongs to a quantitative trait and the complex genetic control of NTSR to three herbicides inhibiting acetylcoenzyme A carboxylase (ACCase) and one inhibiting acetolactate-synthase (ALS) was investigated in black-grass (Petit et al., 2010). Molecular markers and herbicide resistant loci can be achieved through NGS and used for weeds herbicide resistant study.

Conclusion

Economic loss caused by goosegrass resistance to herbicides is an emerging problem in both agriculture and turfgrass. Lack of genomic resource for goosegrass inhibits us to understand of the evolution of herbicide resistance and to identify novel herbicide targets. So here we report the first draft genome of goosegrass. The whole genome sequencing and *de novo* assembly revealed that the final scaffolds were 498 Mb represented 86% of the total goosegrass genome and total of 25,467 genes were predicted in goosegrass genome based on *de novo* method of gene prediction using AUGUSTUS. Furthermore, the gene ontology annotation showed 11,954, 8344, and 14,837 of sequences were involved in biological, cellular and molecular functions, respectively. The genomic resources developed in this project can help us to understand weed biology and weediness, find new herbicide targets, elucidate

targets of know herbicide, understand the mechanisms of evolved herbicide resistance, and find new weed control strategies, which are critical in weed management.

Table 1 Assembly and annotation statistics of goosegrass draft genome

Details	Value
Estimate of genome size	584 Mb
Assembled genome size	497,777,567 bp
Total length of assembled contigs	477,398,776 bp
Number of contigs	34,415
Largest contig	2,426,064 bp
Average contig length	13,872 bp
N50 length (contig)	210,958 bp
N90 length (contig)	27,991 bp
Number of scaffolds	24,063
Total size of assembled scaffolds	497,777,567 bp
N50 length (scaffolds)	233,823bp
N90 length (scaffolds)	31,097 bp
Longest scaffold	2,439,579 bp
Average scaffold length	20,458 bp
GC content	40.83%
Number of genes predicted	25,467

Table 2 Main enzymes in goosegrass genome

Enzyme Commission (EC) classes	Number of sequences
Oxidoreductases	463
Transferases	909
Hydrolases	1,087
Lyases	142
Isomerases	111
Ligases	89

Table 3 Repetitive sequences in goosegrass genome sequence

Repeat type	Number of elements*	Length occupied	Percentage of sequence
Total interspersed repeats		104,378,430 bp	21.20 %
A. Retroelements:	168,497	75,419,902	15.31%
SINEs	4,087	724,658 bp	0.15 %
LINEs	16,584	7,014,468 bp	1.42 %
LTR elements	147,826	67,680,776	13.75 %
B. DNA elements	50,251	14,352,473 bp	2.92 %
C. Unclassified	64,230	14,606,055 bp	2.97 %
Tandem repeats:	75,881	3,865,306 bp	0.78 %
A. Small RNA	4033	701,055 bp	0.14 %
B. Satellites	14	1,371 bp	0.00 %
C. Simples repeats	25,545	1,122,656 bp	0.23 %
D. Low complexity	46,289	2,040,224 bp	0.41 %
Total			21.84 %

* Most repeats fragmented by insertions or deletions have been counted as one element. SINEs, short interspersed elements; LINEs, long interspersed elements; LTR, long terminal repeat.

Table 4 Results of microsatellites and long SSRs search in the assembled goosegrass genome

	SSRs	Long repeat sequences
Total number of sequences examined	24,063	24,063
Total size of examined sequences (bp)	492,285,970	492,285,970
Total number of identified SSRs	115,417	116,940
Number of SSR containing sequences	5,659	6,537
Number of sequences containing more than 1 SSR	3,881	4,229
Number of SSRs present in compound formation	10,076	6,850

Table 5 Anchors and synteny blocks between goosegrass and other five Poaceae species

Species	Anchors	Synteny blocks
Brachypodium vs. Goosegrass	6,645	136
Foxtail millet vs. Goosegrass	8,824	155
Maize vs. Goosegrass	9,269	179
Rice vs. Goosegrass	7,638	156
Sorghum vs. Goosegrass	8,683	148

Table 6 Orthofinder results of seven Poaceae genomes comparation

Percentage of genes in orthogroups	31.4
Number of orthogroups	36,160
Number of species-specific orthogroups	696
Number of genes in species-specific orthogroups	8,919
Percentage of genes in species-specific orthogroups	0.7
Mean orthogroup size	11.1
Number of orthogroups with all species present	10,932
Number of single-copy orthogroups	243

Table 7 Summary of target herbicide resistant genes in goosegrass draft genome

Name	Location	Gene start	Gene end	Length	Exons	Introns
ALS inhibitors	scaffold170_size447559.g14176.t1	22740	24383	1,643	1	0
ACCase inhibitors	scaffold977_size175470.g37610.t1	15495	18636	3,141	4	3
	scaffold726_size237354.g32947.t2	105568	114551	8,983	28	27
HPPD inhibitors	scaffold153_size653311.g13275.t1	383387	385467	2,080	2	1
Dihydropteroate synthase gene	scaffold33_size757691.g4308.t1	368280	369960	1,680	2	1
EPSPS	scaffold1283_size101095.g41869.t1	29872	33187	3,315	8	7
Fatty acid synthase gene	scaffold354_size316591.g22476.t2	136993	145986	8,993	23	22
	scaffold46_size687207.g5455.t1	436142	439204	3,062	11	10
GAPDH (g3p)	scaffold885_size153551.g36057.t1	144698	147519	2,821	12	11
	scaffold33_size757691.g4329.t1	471032	472445	1,413	3	2
GS1	scaffold537_size243955.g28474.t1	176672	187180	10,508	22	21
	scaffold868_size153189.g35714.t1	53667	56665	2,998	10	9
GS2	scaffold197_size765491.g15719.t2	510218	521238	11,020	22	21
	scaffold646_size199344.g31270.t1	177338	181115	3,777	13	12
Phytoene	scaffold3117_size18591.g51177.t1	5114	9211	4,097	14	13
Desaturase gene	scaffold552_size238579.g28909.t1	181768	185961	4,193	13	12
(PDS)	scaffold613_size209948.g30437.t1	59463	64054	4,591	12	11
protox*	scaffold120_size515403.g11237.t1	201980	208022	6,042	17	16
	scaffold163_size476583.g13828.t1	293261	297076	3,815	9	8
psbA	chloroplast genome NC_030486.1	83	1144	1,061	1	0
tubulin alpha	scaffold965_size139305.g37427.t1	123076	125412	2,336	4	3
	scaffold1379_size91218.g42840.t1	61772	64729	2,957	5	4
	scaffold1303_size99324.g42084.t1	49342	54479	5,137	4	3
tubulin beta	scaffold772_size181594.g33901.t1	82682	85434	2,752	3	2
	scaffold8346_size7467.g53640.t1	658	3004	2,346	3	2
ubi	scaffold499_size244782.g27438.t1	205259	208389	3,130	9	8
	scaffold336_size309670.g21905.t1	261301	264312	3,011	5	4
	scaffold84_size607735.g8623.t1	378436	380002	1,566	2	1
	scaffold167_size483343.g14041.t1	168364	171674	3,310	6	5
vlcfa	scaffold118_size525148.g11099.t1	243346	246147	2,801	2	1
	scaffold238_size375850.g17794.t1	308357	309838	1,481	1	0
	scaffold1140_size217858.g40020.t1	62507	63982	1,475	1	0

* For protox, one is in chloroplast and the other is in mitochondria; For other genes, different locations are different isoforms.

Table 8 Summary of number of non-target herbicide resistant genes and family members

Name	Number of members	Number of genes
ABC	158	61
Cyp450	310	85
Glycosyl-transferases	370	152
GS	52	20

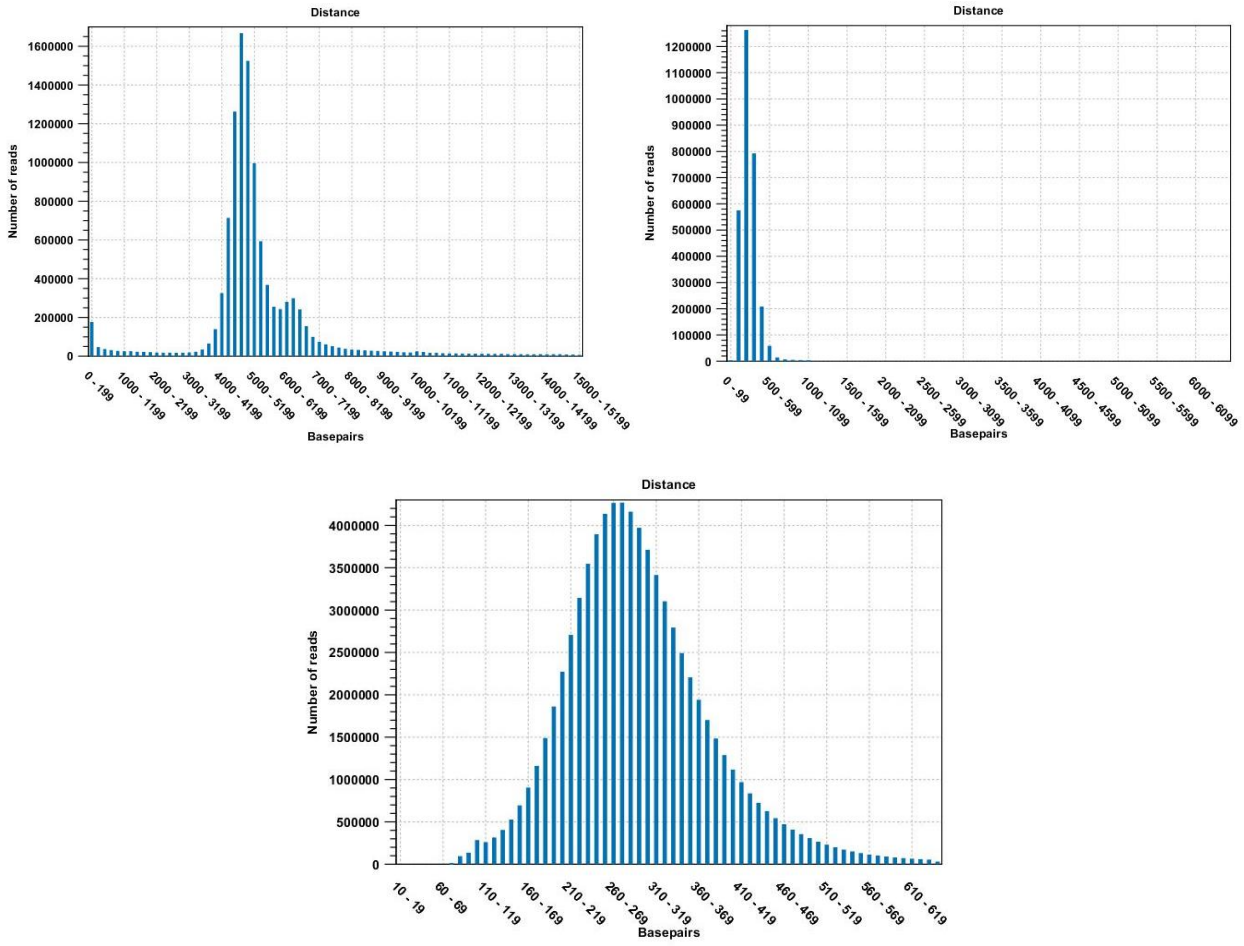


Figure 1 Paired reads distance distribution for all three libraries: (A) Mate-paired sequencing reads (Top left); (B) AU paired-end sequencing reads (Top right); (C) Genewiz paired-end sequencing reads (Bottom).

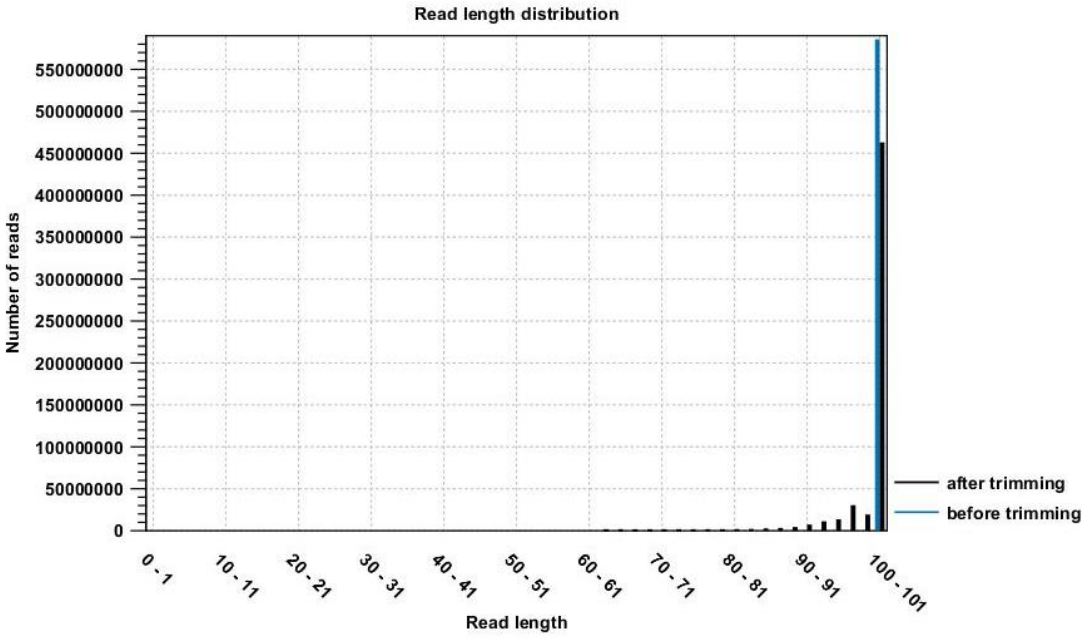


Figure 2 Sequencing reads distribution before and after trimming for three libraries.

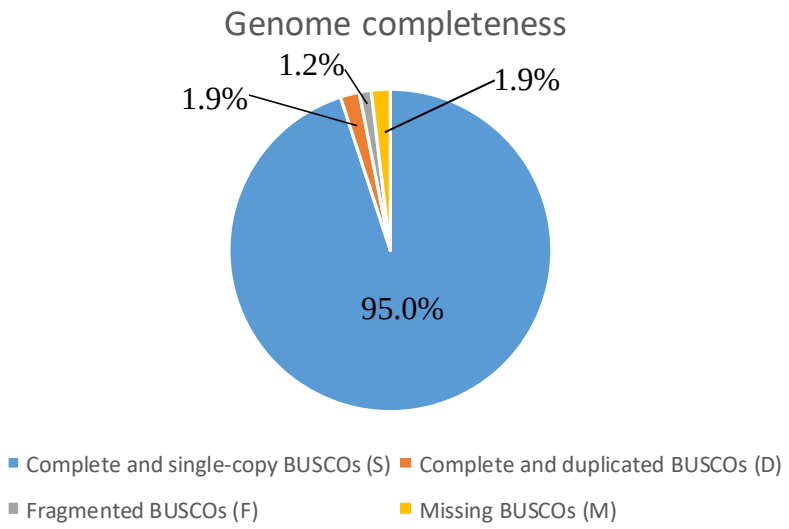


Figure 3 BUSCO results. 95.0% complete single-copy orthologues, 1.9% complete duplicated, 1.2% fragmented and 1.9% missing of orthologues in goosegrass genome.

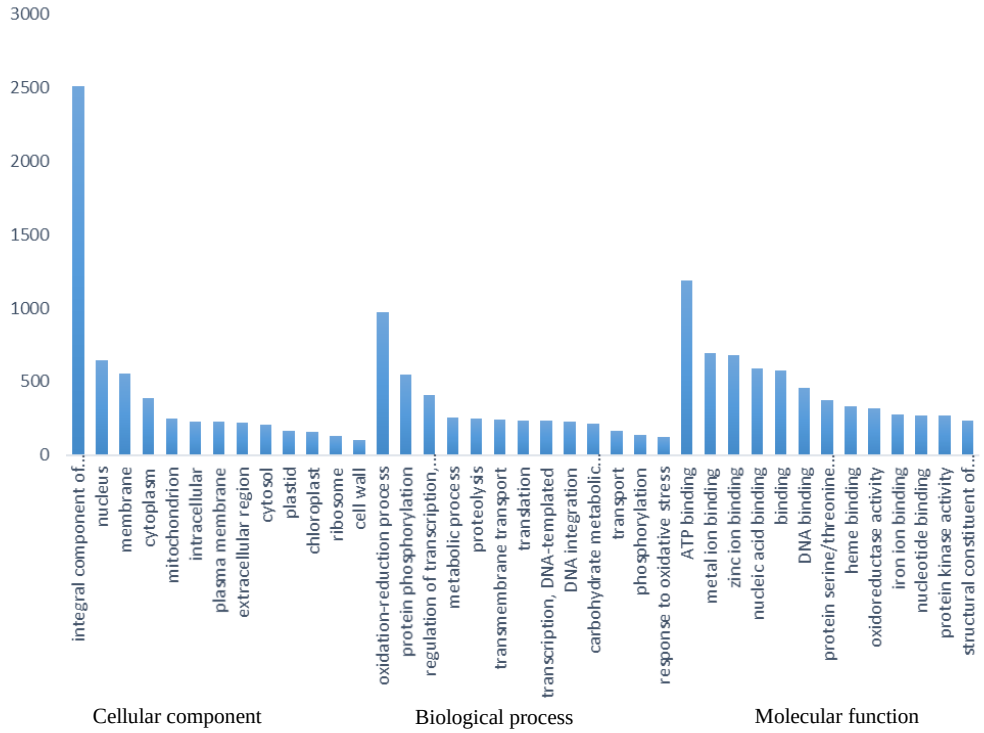
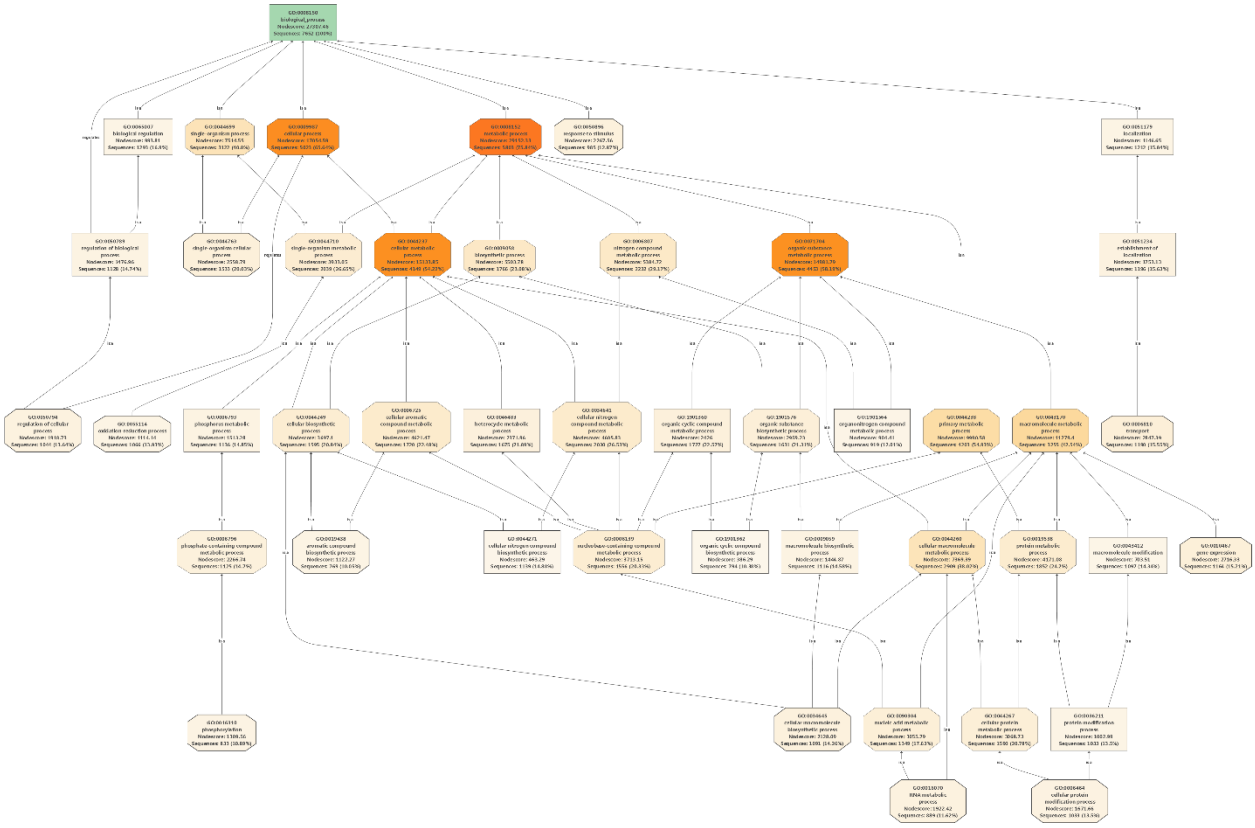
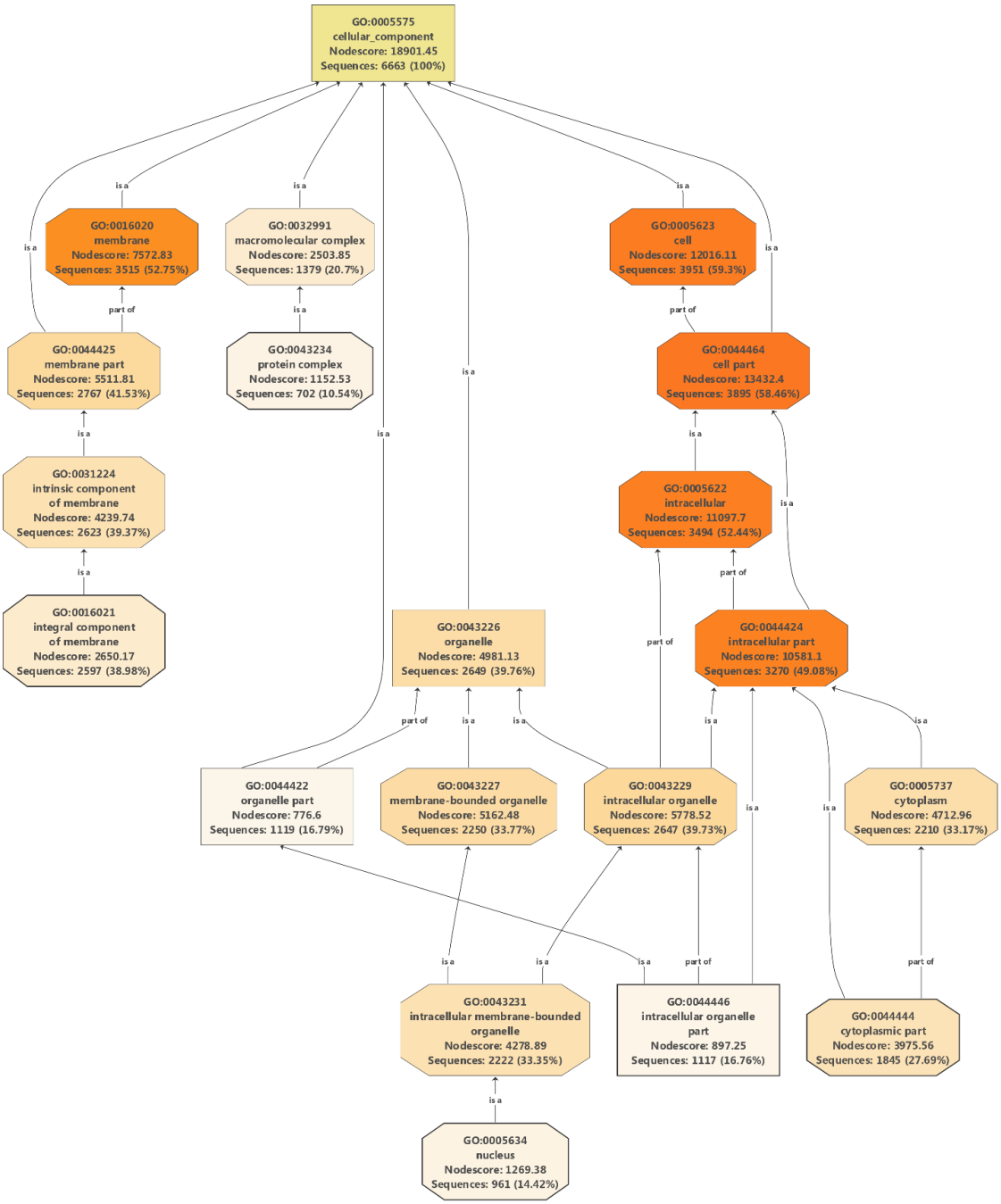


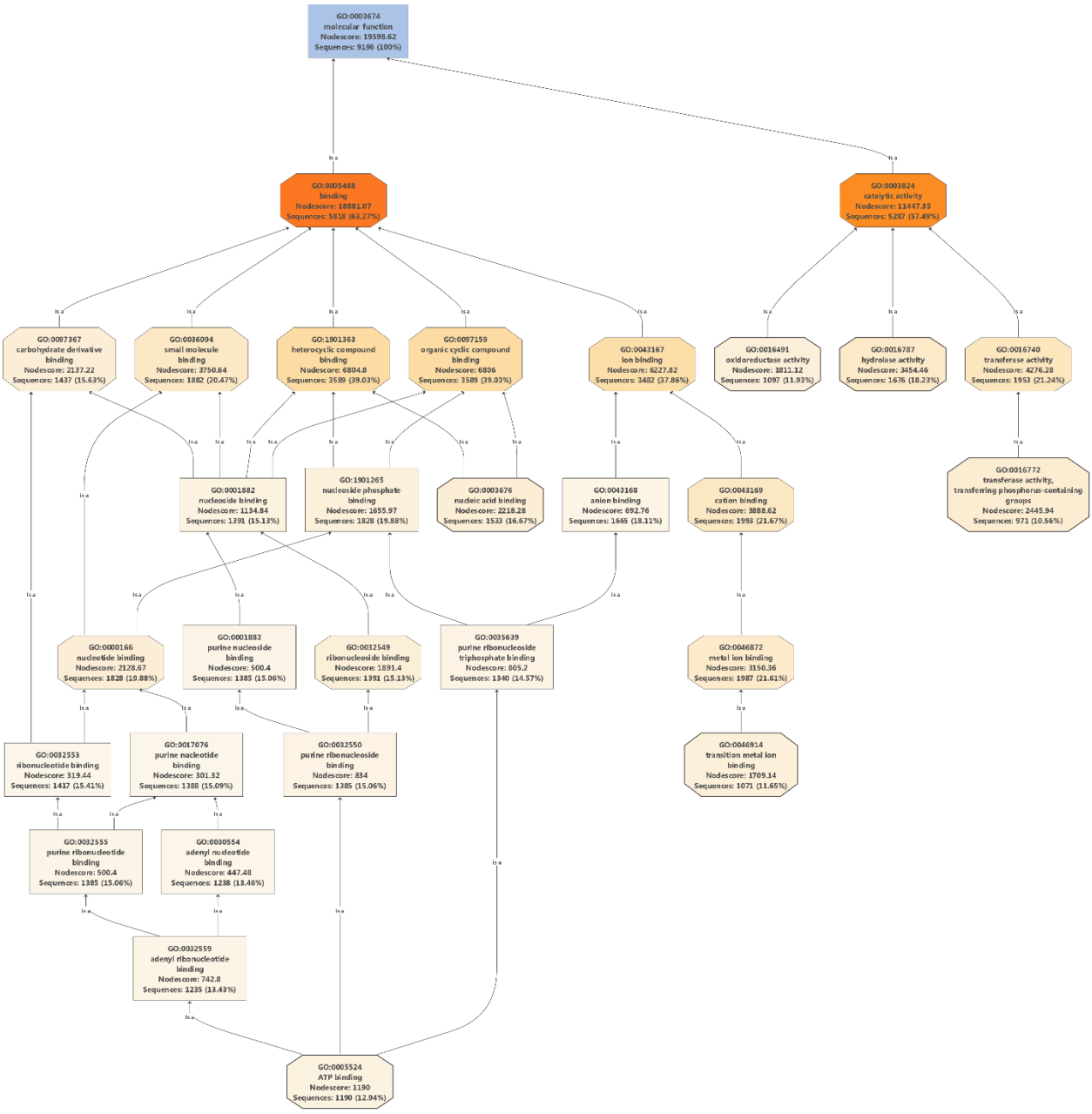
Figure 4 Gene ontology (GO) term categorization and distribution of genes expressed in goosegrass genome. GO-terms were processed using Blast2GO and categorized under cellular component, biological and molecular function GO terms categories, respectively.



(A)

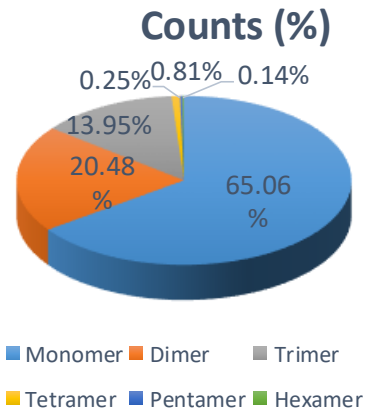

GO:0005575
cellular_component
Nodescore: 18901.45
Sequences: 6663 (100%)
GO:0016020
membrane
Nodescore: 7572.83
Sequences: 3515 (52.75%)
GO:0032991
macromolecular complex
Nodescore: 2503.85
Sequences: 1379 (20.7%)
GO:0005623
cell
Nodescore: 12016.11
Sequences: 3951 (59.3%)
GO:0044425
membrane part
Nodescore: 5511.81
Sequences: 2767 (41.53%)
GO:0043234
protein complex
Nodescore: 1152.53
Sequences: 702 (10.54%)
GO:0044464
cell part
Nodescore: 13432.4
Sequences: 3895 (58.46%)
GO:0031224
intrinsic component
of membrane
Nodescore: 4239.74
Sequences: 2623 (39.37%)
GO:0005622
intracellular
Nodescore: 11097.7
Sequences: 3494 (52.44%)
GO:0016021
integral component
of membrane
Nodescore: 2650.17
Sequences: 2597 (38.98%)
GO:0043226
organelle
Nodescore: 4981.13
Sequences: 2649 (39.76%)
GO:0044424
intracellular part
Nodescore: 10581.1
Sequences: 3270 (49.08%)
GO:0044422
organelle part
Nodescore: 776.6
Sequences: 1119 (16.79%)
GO:0043227
membrane-bounded organelle
Nodescore: 5162.48
Sequences: 2250 (33.77%)
GO:0043229
intracellular organelle
Nodescore: 5778.52
Sequences: 2647 (39.73%)
GO:0005737
cytoplasm
Nodescore: 4712.96
Sequences: 2210 (33.17%)
GO:0043231
intracellular membrane-bounded
organelle
Nodescore: 4278.89
Sequences: 2222 (33.35%)
GO:0044446
intracellular organelle
part
Nodescore: 897.25
Sequences: 1117 (16.76%)
GO:0044444
cytoplasmic part
Nodescore: 3975.56
Sequences: 1845 (27.69%)
GO:0005634
nucleus
Nodescore: 1269.38
Sequences: 961 (14.42%)
is a
part of

(B)

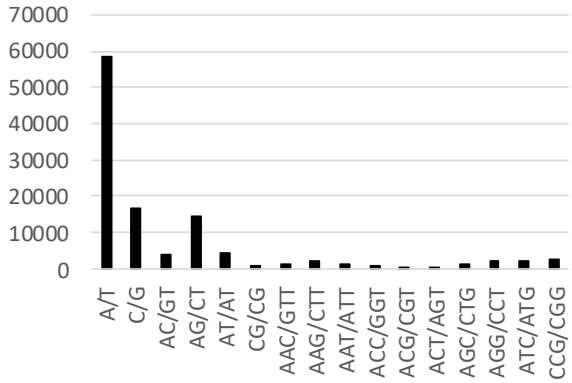


(C)

Figure 5 The relationships of GO terms among each component: (A) Biological process, (B), Cellular component, (C) Molecular function.



(A)



(B)

Figure 6 Summary of SSR markers in the assembled goosegrass genome: (A) Maker types, (B) Numbers of monomer, dimer and trimer makers.

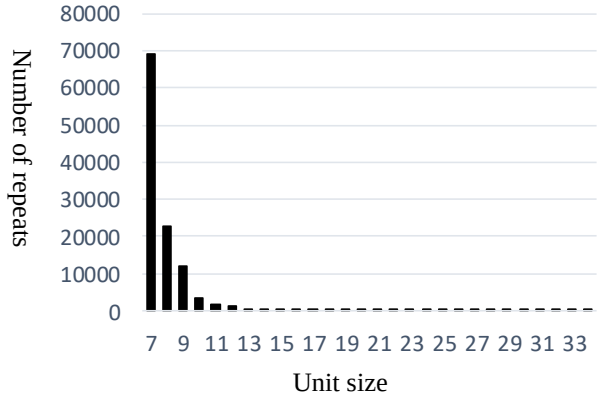


Figure 7 Distribution to different repeat type classes (types of long repeat sequences number great than 100).

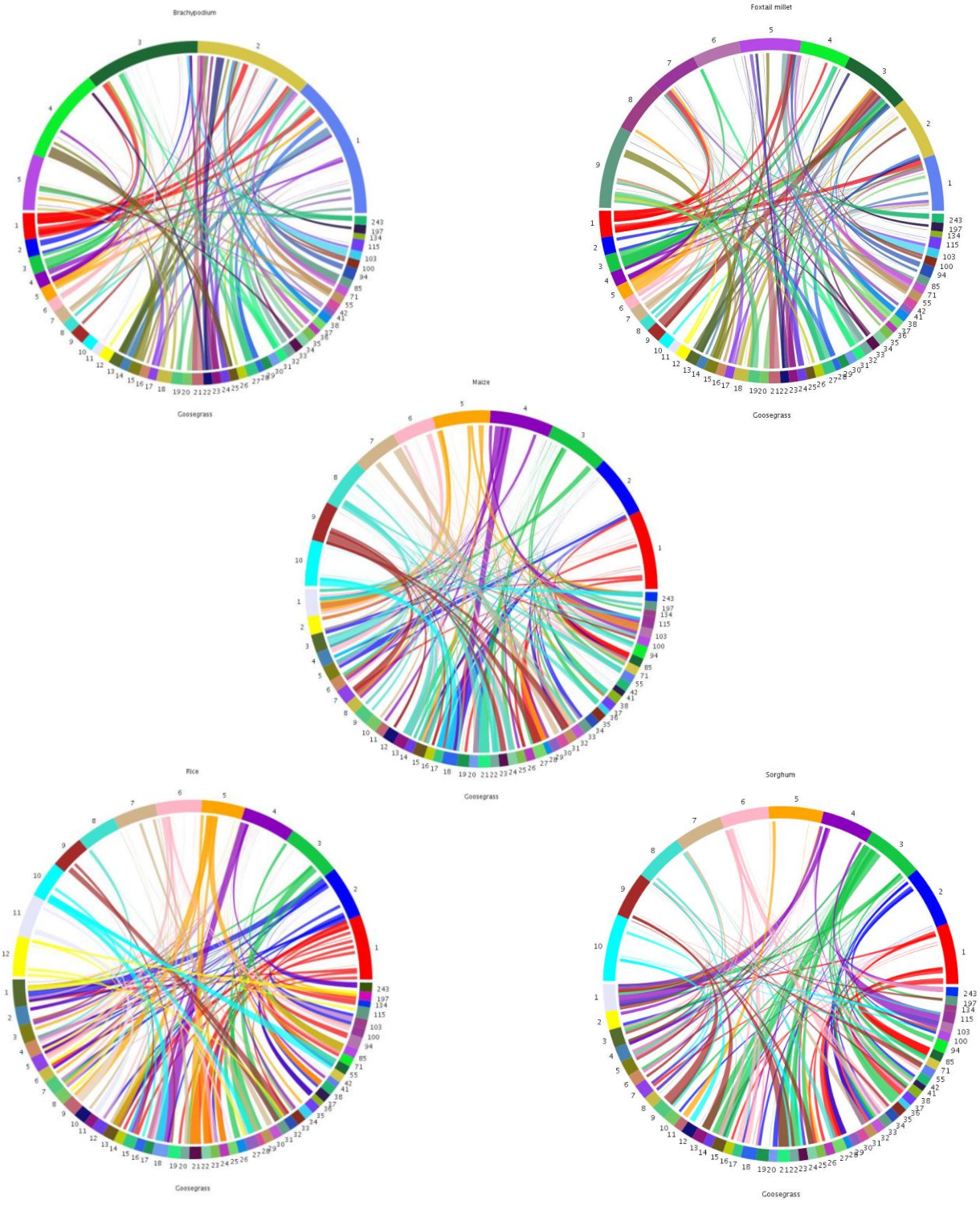


Figure 8 Syntenic genomic blocks of goosegrass with other Poaceae species (Brachypodium, Foxtail millet, Maize, Rice and Sorghum).

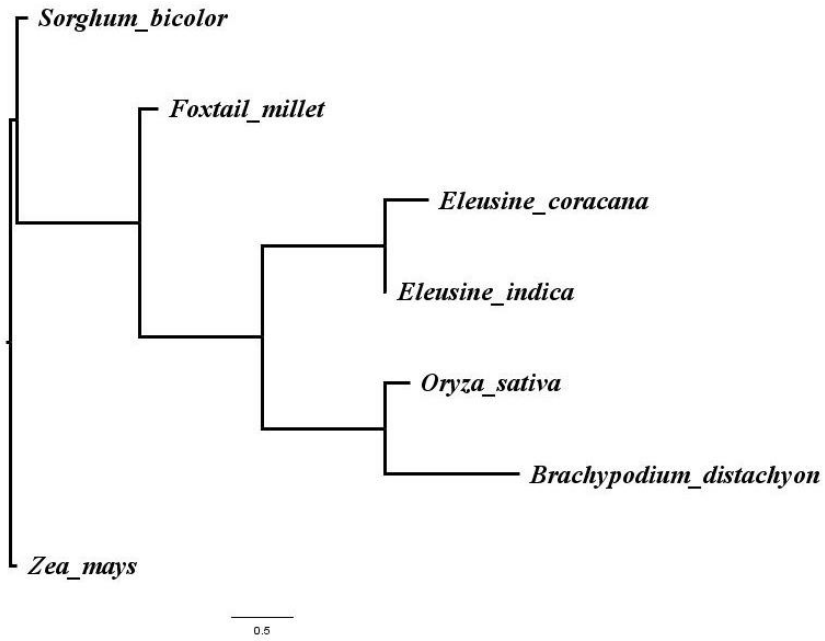


Figure 9 Phylogenetic relationships of six Poaceae species according to the single copy ortholog genes.

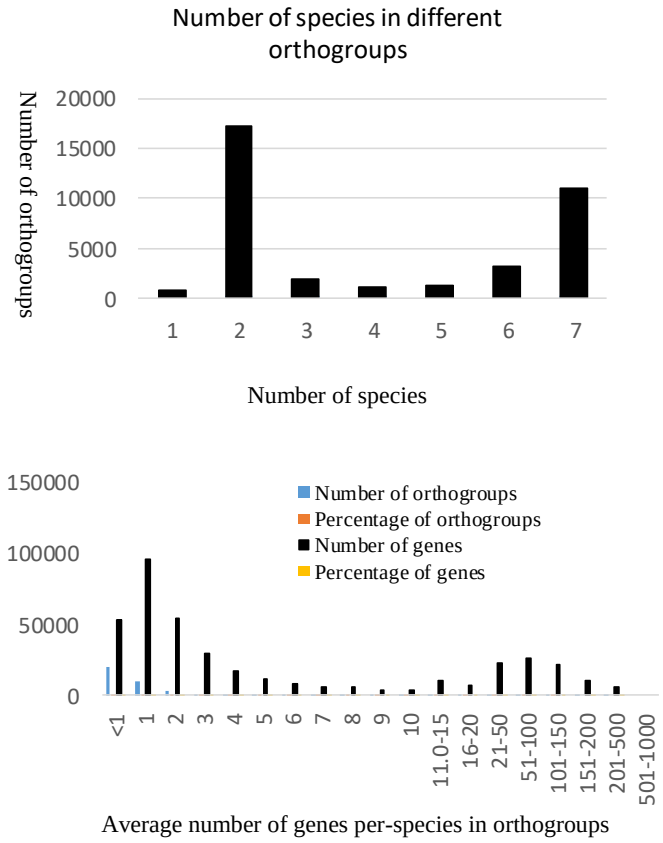


Figure 10 Orthofinder results of seven Poaceae genomes comparation.

Chapter 3. Complete plastid genome sequence of Goosegrass (*Eleusine indica*) and comparison with other Poaceae

Introduction

The chloroplast plays a significant role in numerous plant cell functions, including photosynthesis, the manufacture of certain amino acids and lipids, starch, pigment production, and some key aspects of nitrogen and sulfur metabolism (Cui, 2006). It is considered to have originated from cyanobacteria through endosymbiosis (Raven and Allen, 2003). Chloroplast genomes most commonly exist as a single large circular DNA molecule typically range in size from 120 to 170 kilobase pairs (kb) (Shaw et al., 2007). In angiosperms, chloroplast genomes have a quadripartite organization, composed of two copies of inverted repeat (IR), one large single copy (LSC), and one small single copy (SSC) (Jansen et al., 2005; Zhao et al., 2015). During plant evolution, some chloroplast genes (i.e., *infA*, *rps16*, *ycf1*, *ycf2*, and *ycf4*) were lost through gene transfer to the nucleus or were lost from the cell entirely (Millen et al., 2001). However, the small size and highly conserved of chloroplast genome still makes it suitable and invaluable for complete sequencing and phylogenetic analysis (Cho et al., 2015).

Traditional methods to sequence the chloroplast genome or partial chloroplast genes rely on costly and time-consuming plastid isolation, PCR and amplicon sequencing. With the advent of next-generation sequencing (NGS) technology, new approaches for chloroplast genome sequencing have been gradually proposed due to their high-throughput, time-saving and low-cost (Cronn et al., 2008). The number of available complete chloroplast genomes has increased rapidly due to high-throughput sequencing technology.

Despite the rapidly developing technology, there are relatively few weed genomes and transcriptome available. Transcriptome assemblies have been produced for goosegrass (An et al., 2014; Chen et al., 2015), and no complete chloroplast genome has been reported for any *Eleusine* species. Recently, 78 plastid protein coding loci were sequenced for *E. coracana* (Givnish et al., 2010). Here, we present the

complete chloroplast genome sequence of goosegrass based on a high-throughput sequencing approach and perform comparative analyses of the plastid genomes of goosegrass and other Poaceae.

Materials and methods

Plants and sequencing. The goosegrass population used in this research has been previously utilized for transcriptomic (Chen et al., 2015) and chloroplast genome research (Zhang et al., 2017). The plant seeds were grown in a glasshouse and total genomic DNA was extracted from fresh leaves using DNeasy Plant Mini Kit (Qiagen, CA, USA). Sequencing were conducted at GENEWIZ, Inc. (South Plainfield, NJ, USA) and Auburn University Genomic and Sequencing Lab (http://www.ag.auburn.edu/enpl/gsl/seq.php) using Illumina paired-end and mate-paired genome sequencing technology with reads length 100 bp. In Genewiz, two libraries with insert length 467 bp and 7 kb were prepared according to the manufacturer's standard protocol. Besides, in Genomic and Sequencing Lab at AU, one library with insert length 467 bp was prepared and sequenced (Figure 1). Details see Chapter 1.

Plastid genome assembly. Illumina paired-end data were cleaned with trimmomatic (v0.33; Bolger et al., 2014). Initially reads were mapped against the plastid genomes of *Neyraudia reynaudiana* (NC_024262) and *Setaria italica* (KJ001642) using Bowtie2 (v.2.2.4; Langmead and Salzberg, 2012). Individual reads and their pairs were extracted using samtools (v1.2; Li et al. 2009) and then assembled using Ray (v2.3.1; Boisvert et al., 2010) and PriceTI (v1.0.1; Ruby et al. 2013). Contig maps were examined for cigar strings and depth with high quality contigs scaffolded together using read pairing information and manual alignment. The final genomic assembly was checked for quality and iteratively improved to reflect strict consensus sequence using Bowtie2 (v.2.2.4; Langmead and Salzberg, 2012), samtools (v1.2; Li et al. 2009), Tablet (Milne et al., 2013) and R (v.3.0.2; R Core Team, 2013). Assembly was confirmed by the mapping an additional paired-end library and mate-paired library. Final validation of plastid structure was achieved by using only reads that mapped in their pairs, and in the correct

orientation without the aid of a match bonus in the alignment algorithm, or subsequent soft-clipping of the reads. Besides, we eliminated all reads that bore any difference from the reference and verified that the map maintained complete coverage. Inverted repeats and coding sequences were identified using the NCBI BLAST suit (Acland et al., 2014) and Seaview (Gouy et al., 2010) aligner. GC content was calculated using EMBOSS (v6.4.0.0; Rice et al. 2000). Codon usage was calculated for all exons of protein-coding genes (pseudogenes were not calculated) with Acua 1.0 (Vetrivel et al., 2007).

Annotation. Gene annotation was conducted using DOGMA (Wyman et al., 2004) and manual editing through comparison with the published plastid genome sequence of *Neyraudia reynaudiana* (KF356392). All tRNA genes were further confirmed with tRNA-SE search server (Lowe and Eddy, 1997). The circular chloroplast genome map was drawn using the OGDRAW program (Lohse et al., 2007). Repeat sequences were identified using microsatellite identification tool (Thiel, 2003), and each repeat sequence was ≥ 10 bp. Repeat sequences whose repeating sequence units were arranged from 2-6 bp and repeated not less than three times were considered as SSRs. Repeat sequences with lengths ≥22 bp were considered as long repeat sequences. Two SSRs with interruption lengths each ≤ 100 bp were considered as compound microsatellite repeats.

Comparison to other Poaceae genomes. The software mVISTA (Frazer et al., 2004) was used in Shuffle-LAGAN mode (Frazer et al., 2004) to compare the complete chloroplast genome of *E. indica* to eight representative plastid genomes of other Poaceae: *Anomochloa marantoidea* (GQ329703), *Hordeum vulgare* (NC_008590), *Neyraudia reynaudiana* (KF356392), *Oryza sativa* (KM103382), *Phyllostachys edulis* (HQ337796), *Sorghum bicolor* (NC_008602), *Sporobolus maritimus* (KP176438) and *Zea mays* chloroplast (NC_001666).

All of the protein coding genes in each species' plastome were extracted and concatenated using perl and python scripts. Alignment was did using Mafft and Seaview (Gouy et al., 2010). The concatenated

phylogenetic tree was made in Seaview (Gouy et al., 2010) using Neighbor Joining method with replicates 1000.

Results and discussion

Plastome sequencing and assembly. Using the Illumina sequencing technology, we obtained 220,526,422 (mate-paired) and 200,555,780 (paired-end) raw reads of 100 bp in length from Genewiz and 164,677,932 raw reads of 100 bp from the Auburn Sequencing Center. Illumina adaptors and barcodes were removed from raw reads. The numbers of reads after trimming for quality were 219,046,803, 199,605,202 and 164,598,344 respectively. For paired-end data, the chloroplast reads were 16, 408,090 with average coverage 9496.304, maximum coverage 20,985, average fragment size 279.8 and maximum fragment size 518.0; For mate-paired data, the chloroplast reads were 1,319,964 with average coverage 979.1, maximum coverage 1,937.0, average fragment size 4889 and maximum fragment size 9919. So these data is more than enough to assemble the plastid genome. The length for the *E. indica* whole chloroplast genome sequence was 135,151 bp. 9.6 M reads mapped to the finished plastid genome and of these 8.8 M reads mapped with no insertions, deletions or mismatches of any kind. Additional read sets mapped to the finished genome validated the assembly. In this plastome sequence, the total length for LSC, SSC and IR regions were 80,667 bp, 12,646 bp and 20,919 bp, respectively. The annotated genome sequence has been submitted to GenBank (accession number KU833246).

Plastome organization and gene content. In the *Eleusine indica* plastid genome, 108 unique genes were identified, including 76 protein-coding genes, 28 tRNA genes, and 4 rRNA genes (Table1, Figure1). The tRNA-coding genes represent all 20 amino acids and were distributed throughout the entire genome, one in the SSC region, 19 in the LSC region and eight in the IR region. Four rRNA genes were also identified in this plastome, completely duplicated in the IR regions. In total, eight genes coding for tRNA, four rRNA genes and six protein-coding genes (*rps19, rpl2, rpl23, ndhB, rps7, rps12*) were completely

duplicated in the IR regions. Therefore, the total number of genes present in the goosegrass chloroplast genome is 126 (Figure 1). Sequence analysis indicates 44.01% of the genome sequences encoding proteins, 2.11% encoding tRNAs, and 6.80% encoding rRNAs, whereas the remaining 16.71% are introns and 30.38% are other non-coding regions which include pseudogenes and intergenic spacers.

The GC content of the *E. indica* chloroplast genome is 38.19%, which is consistent with reported Poaceae (rice, 38.95%; maize, 38.5%) (Maier et al., 1995; Wu and Ge, 2014). The GC content of the LSC and SSC region are 36.09% and 32.33%, respectively, whereas that of the IR region is 44.01%. The high GC content in the IR regions is due to the reduced presence of AT nucleotides in the duplicate rRNA genes: *rrn16*, *rrn23*, *rrn4.5*, and *rrn5*.

In the goosegrass chloroplast genome, there are 15 intron-containing genes (Table 1). Among them, 14 genes (nine protein-coding and five tRNA genes) have a single intron and one gene (*ycf3*) has two introns. Except for *rps12*, among the 14 genes with introns, 9 (six protein-coding and three *tRNA* genes) are located in the LSC, one protein coding in the SSC and 4 (two protein coding and two *tRNAs*) in the IR region. The *rps12* gene is a trans-spliced gene: its 5' end exon is located in the LSC region and the two remaining exons are located in the IR regions. The *trnK-UUU* has the largest intron (2, 483 bp) which contains an entire additional gene, *matK*.

In addition, there are 59,475 nt and 19,825 codons in the goosegrass plastome in total, which represent the coding regions of 76 protein-coding genes. According to the sequences of protein-coding genes and tRNA genes, the frequency of codon usage was obtained (Table 2). Among these codons, 2151 codons for Leucine (equal to 10.85% of the total) and 214 codons for cysteine (equal to 1.08% of the total), which are the most and the least abundant amino acids, respectively. The codon usage is biased towards a high representation of A and T at the third codon position, which is similar to the majority of angiosperm plastid genomes (Tangphatsornruang et al., 2010; Qian et al., 2013; Curci et al., 2015).

Simple sequence repeats in the E. indica plastome. Simple sequence repeats (SSRs), known as microsatellites and short tandem repeats (STRs) are valuable genetic molecular markers for population

genetics (He et al., 2012) and phylogenetic analysis (Melotto-Passarin et al., 2011; Nie et al., 2012). Using the microsatellite identification tool (MISA), 37 SSRs were found in the *Eleusine indica* chloroplast genome (Table 3). Among these SSRs, there are 27 homopolymers, 3 dipolymers, 1 tripolymer, and 6 tetrapolymers. In 30 homopolymers and dipolymers, 27 SSRs only contain A or T bases. In the other 7 SSRs, more than half of the bases contain A or T bases. So SSRs in *E. indica* chloroplast genomes are AT rich. Similar results have been reported in Poaceae (Melotto-Passarin et al., 2011) and other families (Yi and Kim, 2012; Martin et al., 2013). The information from these SSRs will provide useful sources for developing primers and studying specific SSR loci in population samples.

Except for SSRs, repeats with lengths ≥ 22 bp were considered as long repeat sequences in the goosegrass plastome, and 8 long repeat sequences were detected (Table 4). Only one, within the *rpoC2* gene, was located in a coding region. Compound microsatellites are a special variation of microsatellites in which two or more individual microsatellites are found directly adjacent to each other (Kofler et al., 2008). Two SSRs with interruption lengths each ≤ 100 bp were considered as compound microsatellite repeats in our study. SSRs in eukaryotic and prokaryotic genomes have been well documented, however, compound microsatellites are still rarely reported although they can provide insight into the evolution of microsatellites (Bull et al., 1999; Chen et al., 2011). In the *E. indica* plastid genome, five compound microsatellites were detected (Table 5). Among these compound microsatellites, 3 were located in intergenic regions, 2 in protein coding genes (*infA* and *ndhH*).

Comparison to other Poaceae plastid genomes. In size, *E. indica* has the second smallest of nine reference Poaceae chloroplast genomes. It is around 5.6 kb smaller than *Sorghum bicolor* and 0.6 kb bigger than *Oryza sativa* plastome (Table 6). For the phylogenetic relationship among 9 Poaceae species (Figure 4), *E. indica* is closer to *Sporobolus maritimus* (97.02%) and then *Neyraudia reynaudiana* (95.06%), which is consistent with percent identity in Table 6.

Sequence identity comparisons among the nine Poaceae chloroplast genomes were made with mVISTA with the annotated *E. indica* sequence as a reference (Figure 2). Although some divergent

regions are apparent, the aligned sequences indicate a fairly conservative pattern of evolution. Of all genes, *ndhF* is the most divergent (average pairwise divergence: 6.46%) and *rpoC2* also show high divergence (average pairwise divergence: 5.15%). Noncoding regions show higher sequence divergence among nine chloroplast genomes, with the *trnH-GUG-psbA, rps16-trnQ-UUG, petA-psbJ, petN-trnC-GCA, trnC-GCA-rpoB, psbE-petL, ndhF-rpl32* and *psbM*/*trnD-GUC* having the highest levels of divergence (Figure 2).

The expansion and contraction of the inverted repeat IR regions and the single copy boundary regions can result in length variation of plastid genomes overall. In fact, the locations of the LSC/IR and SSC/IR junctions are usually considered as an index of chloroplast genome evolution (Zhang et al., 2013). The IR/SSC borders, IR/LSC borders, and the adjacent genes, were compared across the 9 Poaceae chloroplast genomes (Figure 3). No genes overlap the LSC/IR borders. In *S. bicolor* and *H. vulgare*, the IR regions extended to 2 bp beyond the trnH and the LSC regions extended to 53 bp beyond IR regions. Except for *A. marantoidea*, six other species show similar LSC/IR borders, noticeably different than plastids from non-Poaceae families (Nie et al., 2012; Curci et al., 2015).

The *ndhF* gene was entirely located in the SSC region in *P. edulis, A. marantoidea,* and *H. vulgare*, but varied in distance from the SSC/IRa border with length 127 bp, 105 bp, and 61 bp, respectively. However, this gene extended across IRa and SSC to varying degrees, such as *S. bicolor*, *Z. mays, S. maritimus*, *N. reynaudiana*, and *E. indica*. This gene extended to 9 bp beyond SSC/IRa border in *O. sativa* species. The *ndhH* gene in *E. indica* overlaps the SSC/IRa border by 4 bp, and that of *P. edulis, A. marantoidea, H. vulgare,* and *O. sativa* by 186 bp, 178 bp, 216 bp, and 162 bp, respectively. However, in *S. bicolor*, *Z. mays*, *S. maritimus*, *N. reynaudiana*, the *ndhH* gene is entirely located in the SSC region.

Conclusion

In this study, we assembled and analyzed the complete nucleotide sequence of the plastid genome of *Eleusine indica* using Illumina high-throughput sequencing technology. Compared to eight other representative plastid genomes from Poaceae, this genome has a relatively small size, but the organization and gene content is highly similar. The discovery of tandem repeats in the chloroplast genome of *E. indica* will provide useful information for future phylogenetic and population genetics study in this genus.

Table 1 Plastid genome gene contents in *Eleusine indica*

Category	Group	Gene name
Photosynthesis	Subunits of NADH-dehydrogenase	ndhA†, ndhB†, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
	Subunits of photosystem I	psaA, psaB, psaC, psaI, psaJ
	Subunits of photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
	Subunits of cytochrome b/f complex	petA, petB†, petD†, petG, petL, petN
	Subunits of ATP synthase	atpA, atpB, atpE, atpF†, atpH, atpI
	Large subunit of rubisco	rbcL
Replication	rRNA genes	rrn4.5, rrn5, rrn16, rrn23
	tRNA genes	trnA-UGC†, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-UCC, trnH-GUG, trnI-CAU, trnI-GAU†, trnK-UUU†, trnL-CAA, trnL-UAG, trnL-UAA†, trnM-CAU, trnN-GUU, trnP-UGG, trnQ-UUG, trnR-UCU, trnR-ACG, trnS-GGA, trnS-GCU, trnS-UGA, trnT-UGU, trnV-GAC, trnV-UAC†, trnW-CCA, trnY-GUA
	Small subunit of ribosome	rps2, rps3, rps4, rps7, rps8, rps11, rps12†, rps14, rps15, rps16†, rps18, rps19
	Large subunit of ribosome	rpl2†, rpl14, rpl16†, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36
	DNA dependent RNA polymerase	rpoA, rpoB, rpoC1, rpoC2
Other	Translational initiation factor	infA
	Maturase	matK
	Protease	clpP
	Envelope membrane protein	cemA
	c-type cytochrome synthesis gene	ccsA
Genes of unknown function	Open Reading Frames (ORF, ycf)	ycf4, ycf3†

Note: † indicates genes containing one or more introns.

Table 2 Codon-anticodon recognition pattern and codon usage for *Eleusine indica* plastid genome

AmAcid	Codon	Number	/1000	Fraction	tRNA	AmAcid	Codon	Number	/1000	Fraction	tRNA
Ala	GCG	137.00	6.91	0.11	trnA-UGC	Pro	CCG	85.00	4.29	0.10	
Ala	GCA	372.00	18.76	0.30		Pro	CCA	233.00	11.75	0.27	
Ala	GCU	565.00	28.50	0.45		Pro	CCU	349.00	17.60	0.41	trnP-UGG
Ala	GCC	175.00	8.83	0.14		Pro	CCC	192.00	9.68	0.22	
Cys	UGU	162.00	8.17	0.76	trnC-GCA	Gln	CAG	143.00	7.21	0.21	
Cys	UGC	52.00	2.62	0.24		Gln	CAA	529.00	26.68	0.79	trnQ-UUG
Asp	GAU	560.00	28.25	0.79	trnD-GUC	Arg	AGG	106.00	5.35	0.09	
Asp	GAC	147.00	7.41	0.21		Arg	AGA	364.00	18.36	0.30	trnR-UCU
Glu	GAG	254.00	12.81	0.24		Arg	CGG	92.00	4.64	0.08	
Glu	GAA	783.00	39.50	0.76	trnE-UUC	Arg	CGA	264.00	13.32	0.22	
Phe	UUU	717.00	36.17	0.65	trnF-GAA	Arg	CGU	292.00	14.73	0.24	trnR-ACG
Phe	UUC	388.00	19.57	0.35		Arg	CGC	98.00	4.94	0.08	
Gly	GGG	270.00	13.62	0.18	trnG-UCC	Ser	AGU	290.00	14.63	0.21	trnS-UGA
Gly	GGA	585.00	29.51	0.40		Ser	AGC	106.00	5.35	0.08	
Gly	GGU	469.00	23.66	0.32		Ser	UCG	110.00	5.55	0.08	
Gly	GGC	148.00	7.47	0.10		Ser	UCA	237.00	11.95	0.17	
His	CAU	333.00	16.80	0.74	trnH-GUG	Ser	UCU	385.00	19.42	0.28	
His	CAC	118.00	5.95	0.26		Ser	UCC	266.00	13.42	0.19	trnS-GCU
Ile	AUA	511.00	25.78	0.31	trnI-CAU	Thr	ACG	121.00	6.10	0.11	
Ile	AUU	831.00	41.92	0.51	trnI-GAU	Thr	ACA	314.00	15.84	0.30	trnT-UGU
Ile	AUC	303.00	15.28	0.18		Thr	ACU	434.00	21.89	0.41	trnT-GGU
Lys	AAG	277.00	13.97	0.27		Thr	ACC	195.00	9.84	0.18	
Lys	AAA	740.00	37.33	0.73	trnK-UUU	Val	GUG	162.00	8.17	0.14	
Leu	UUG	389.00	19.62	0.18	trnL-CAA	Val	GUA	439.00	22.14	0.38	trnV-UAC
Leu	UUA	732.00	36.92	0.34	trnL-UAA	Val	GUU	433.00	21.84	0.37	trnV-GAC
Leu	CUG	120.00	6.05	0.06	trnL-UAG	Val	GUC	124.00	6.25	0.11	
Leu	CUA	312.00	15.74	0.15		Trp	UGG	345.00	17.40	1.00	trnW-CCA
Leu	CUU	463.00	23.35	0.22		Tyr	UAU	568.00	28.65	0.79	trnY-GUA
Leu	CUC	135.00	6.81	0.06		Tyr	UAC	148.00	7.47	0.21	
Met	AUG	465.00	23.46	1.00	trnM-CAU	End	UGA	20.00	1.01	0.21	
					trnfM-CAU	End	UAG	21.00	1.06	0.22	
Asn	AAU	585.00	29.51	0.74	trnN-GUU	End	UAA	53.00	2.67	0.56	
Asn	AAC	209.00	10.54	0.26							

Note: /1000: Relative frequency for a specific codon in 1000 codons. Fraction, frequency of codon usage for a specific amino acid.

Table 3 Simple sequence repeats in the *Eleusine indica* plastid genome

SSR type	Unit	length	No.SSRs	Position on genome
P1	A	10	5	3539-3548, 12353-12362, 30326-30335, 45062-45071, 63650-63659
		11	7	12664-12674, 29879-29889, 36787-36797, 47298-47308, 48421-48431, 58953-58963, 71629-71639
		13	2	44485-44497, 51085-51097
P1	T	10	4	8750-8759, 11619-11628, 42145-42154, 42322-42331
		11	2	14961-14971, 104281-104291
		12	2	32748-32759, 50257-50268
		13	2	7727-7739, 80130-80142
P1	G	10	1	93953-93962
		12	1	47137-47148
P1	C	10	1	121857-121866
P2	AT	10	2	26214-26223, 51554-51563
P2	TA	12	1	64721-64732
P3	TTC	12	1	80577-80588
P4	AGAA	12	1	68582-68593
P4	TTCT	12	1	43513-43524
P4	GTAG	16	1	52486-52501
P4	AACG	12	1	98619-98630
P4	AATA	12	1	106912-106923
P4	TCGT	12	1	117188-117199

Note: Pn means repeat unit including n bases.

Table 4 Long repeat sequence in the *Eleusine indica* plastid genome

Repeat pattern	Size (bp)	Position	Location
(TATATTTTTTT)2	22	71895-71916	Intergenic region
(TAGTAGTCTTA)2	22	101492-101513	Intergenic region
(TAAGACTACTA)2	22	114306-114327	Intergenic region
(TTCAAAACACATA)2	26	6592-6617	Intergenic region
(ATGATATAAAATCGAA)2	32	42513-42544	Intergenic region
(GAAGAAATATGGATAAAAG)2	38	5682-5719	Intergenic region
(TTTTTCTTGTGTCGATTCTT)2	40	61087-61126	Intergenic region
(ATATAGGACCCTAGAGGAAGA)2	42	27370-27411	rpoC2

Table 5 Compound microsatellite repeats in the *Eleusine indica* plastid genome

Compound SSR	Size (bp)	Position	Location
(ATAC)3ataattgtatgtataactataagaaaaagggagggaaattggataagaaagattcttttctatac(AT)6	89	16750-16838	Intergenic region
(TA)5gtatatgaatcaaataatatatggaccaaagaaagactacttcttctggatccaaaattaataaaataaagaaatcca(T)11	99	65720-65818	Intergenic region
(T)10ctctccta(T)10	28	76666-76693	infA
(A)11gattgaatcagtttactttctattcctattc(T)10	52	104032-104083	Intergenic region
(ATCC)3ataaggtagatcggcagctactcctccaatgcgaaagtaattatgcatcattcgcatacctgtagcagcttcaaatagatcatatatcaat(TC)5	113	113718-113830	ndhH

Table 6 Size comparison among nine completely sequenced Poaceae plastomes

Species	Accession Number	Genome Size (bp)	LSC (bp)	SSC (bp)	IR (bp)	Identidty (%)
Sorghum bicolor	NC_008602	140754	83733	12503	22259	93.10
Zea mays	NC_001666	140384	82352	12536	22748	91.14
Phyllostachys edulis	HQ337796	139679	83213	12870	21798	95.04
Anomochloa marantoidea	GQ329703	138412	82274	12162	21988	85.54
Hordeum vulgare	NC_008590	136462	81671	12701	21045	88.32
Sporobolus maritimus	KP176438	135592	80858	12714	21010	97.02
Neyraudia reynaudiana	KF356392	135367	80616	12695	21028	95.06
Eleusine indica	KU833246	135151	80667	12646	20919	1
Oryza sativa	KM103382	134551	80604	12343	20802	87.40

Note: Species are ordered by genome size. LSC: Large Single-Copy SSC: Small Single-Copy IR: Inverted Repeat

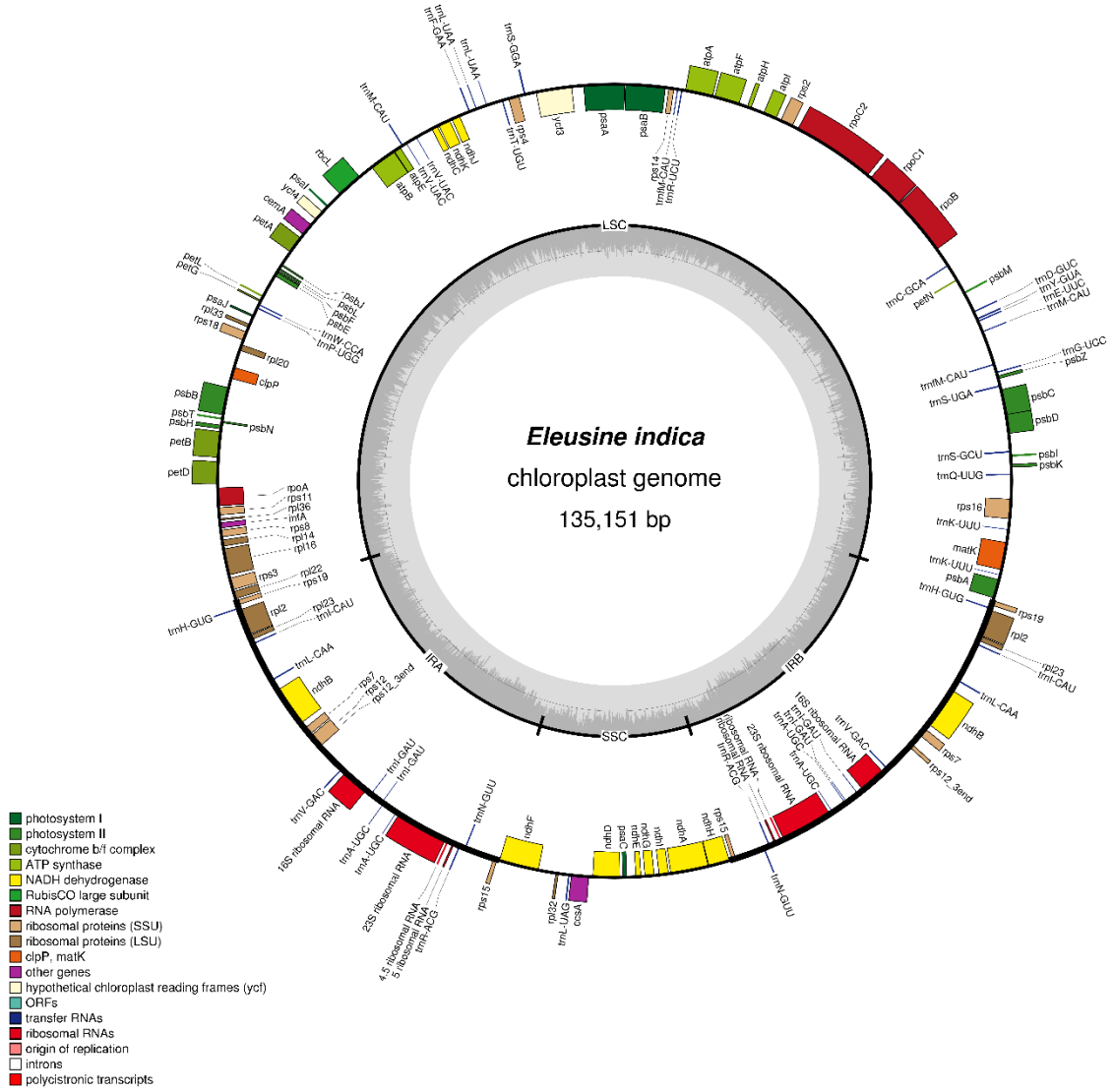


Figure 1 Gene map of the *Eleusine indica* plastid genome sequence. Genes shown outside the outer circle are transcribed counterclockwise, and those inside are transcribed clockwise. Genes belonging to different functional groups are color coded. The innermost darker gray corresponds to GC while the lighter gray corresponds to AT content.

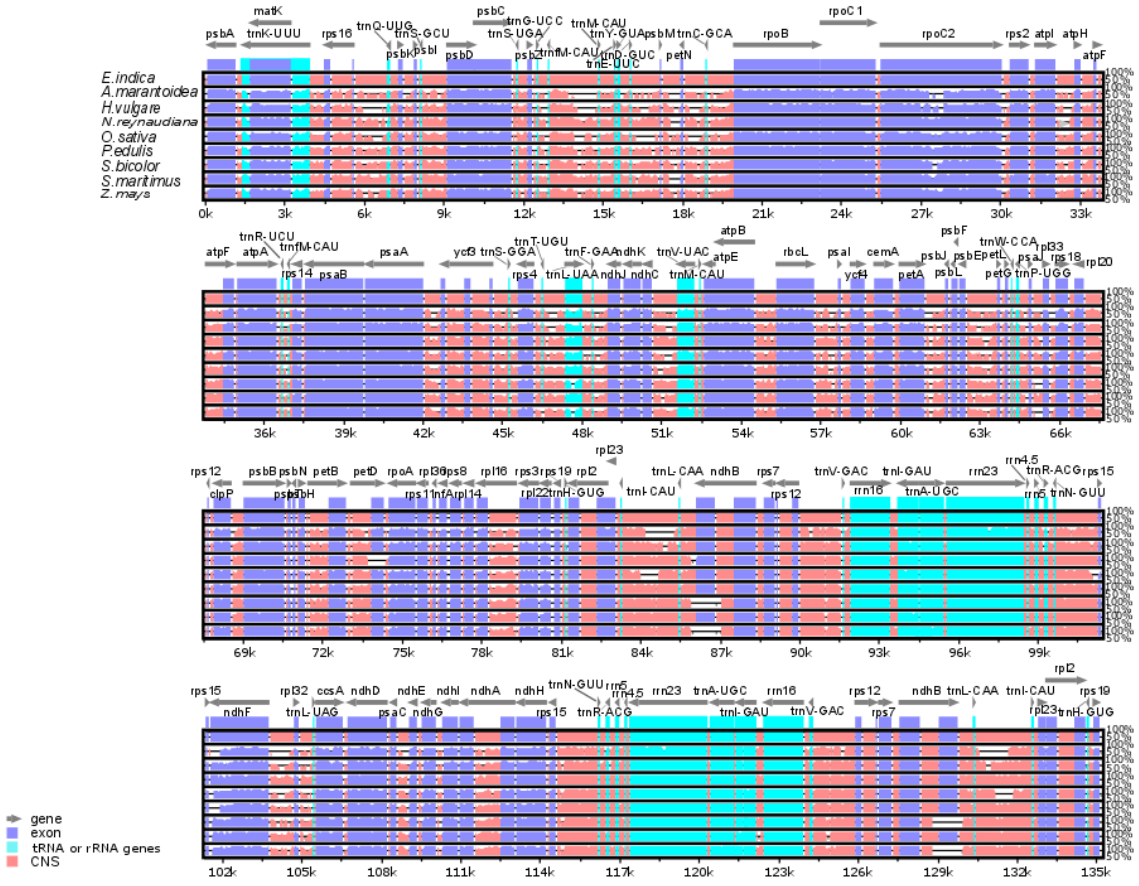


Figure 2 Percent identity plot among the plastid genomes of nine species of Poaceae, using *Eleusine indica* as a reference. Vertical scale indicates the percent identity, ranging from 50% to 100%. The horizontal axis indicates the coordinates within the *Eleusine* genome. Arrows indicate annotated genes and their transcriptional direction.

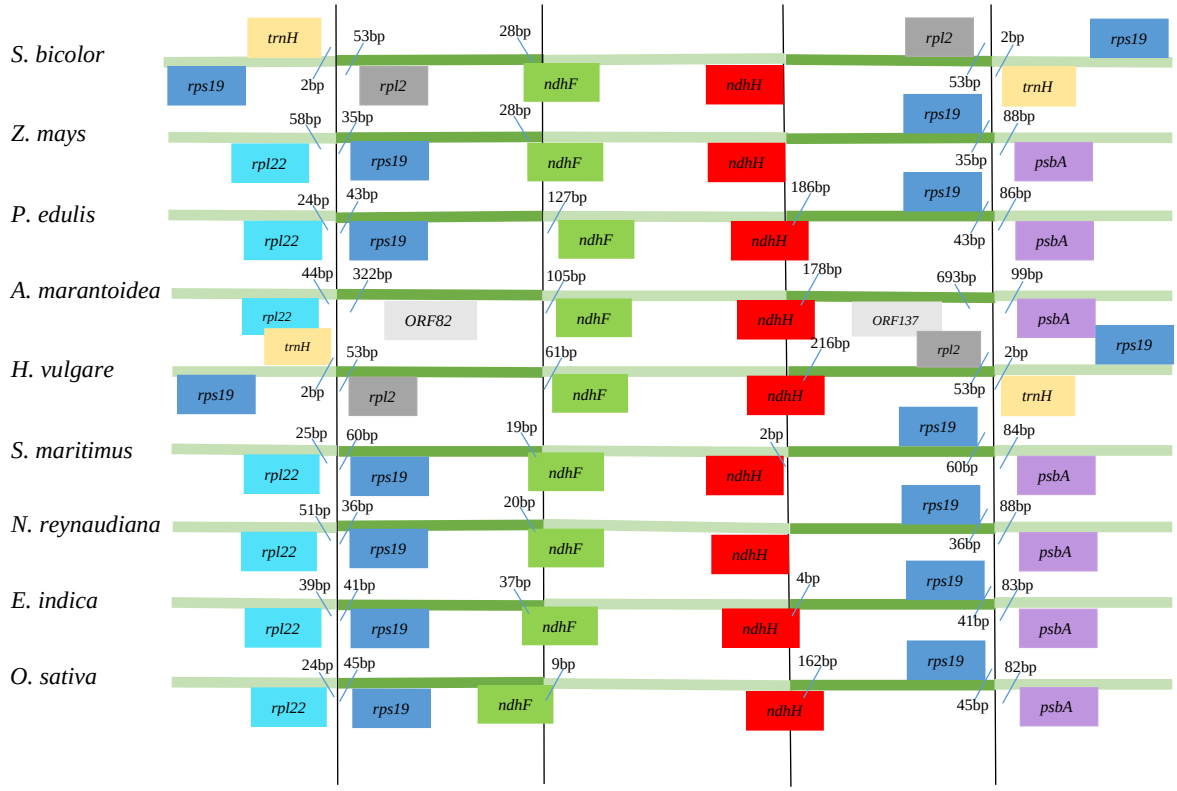


Figure 3 Comparison of the borders of LSC, SSC, and IR regions among nine sequenced Poaceae chloroplast genomes. Genes above lines are transcribed forward while genes below the lines are transcribed reversely.

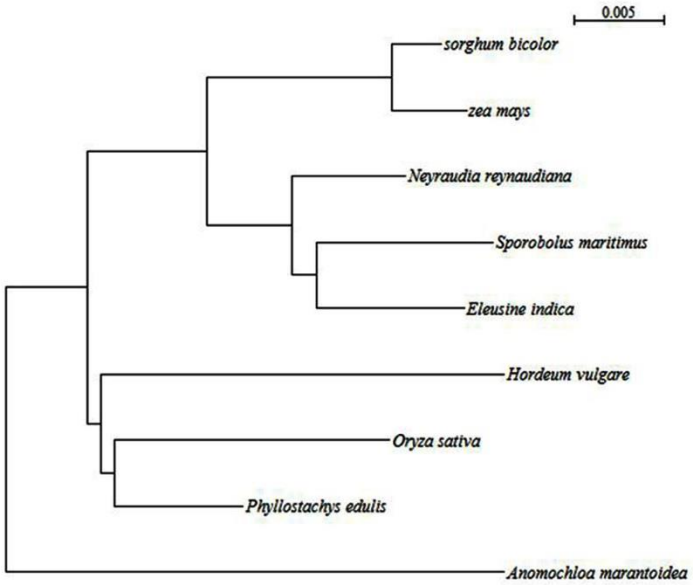


Figure 4 The concatenated phylogenetic tree is based on 76 protein-coding genes using distance method. The bootstrap value is 100%.

Chapter 4. Constructing *Eleusine* transcriptome references for determination of finger millet (*Eleusine coracana*) heritage

Introduction

Despite the rapidly developing technology, there are still relatively few assembled genomes or transcriptomes available for orphan crops. The major limitation to develop orphan crops is that information on germplasm is not readily accessible, found outside of traditional peer-reviewed academic publishing, or written in languages not well-known to the scientific community concerned (Hammer and Heller, 1998). In addition, existing knowledge on the genetic potential of minor crops is limited with few genetic resources, like genomes, transcriptomes and ESTs available online compared to major or industrial crops (Dawson et al., 2009). Further, lack of information about its origin and ancestry slows breeding of minor crops. In plant breeding, paternal and maternal germplasm with desirable traits are collected and desirable traits introduced to the cultivated species through hybridization and backcrossing. Knowing the parentage aided the development of peanut since wild diploid Arachis species possess genetic variability in pest and disease resistance traits, which could be used to improve the cultivated peanut (Chopra et al., 2016; Stalker and Moss, 1987). Therefore, it is beneficial to know the origin and ancestry of a crop to improve breeding.

Transcriptome assemblies have been produced for *E. indica* (An et al., 2014; Chen et al., 2015a) and *E. coracana* (Kumar et al., 2014; Rahman et al., 2014), and 78 plastid protein coding loci were sequenced for *E. coracana* (Givnish et al., 2010). Recently, complete chloroplast genome has been reported for *E. indica* (Zhang et al., 2017) and draft nuclear genome has been reported for *E. coracana* (Hittalmani et al., 2017; Hatakeyama et al., 2017), and Hatakeyama et al. (2017) used a novel multiple hybrid assembly workflow which is suitable for the assembly of complex allotetraploid species. Although there are more and more genomic resources of *E. coracana*, it is still hard to improve this species through plant breeding

since people don't know its evolutionary origin. The maternal genome donor has long been established to be *Eleusine indica*, however the paternal genome donor remains elusive. *E. indica*, an annual diploid (2n = 2x = 18), is most commonly mentioned as the maternal genome donor based on genomic in situ hybridization (Bisht and Mukai, 2001a; Hilu, 1988 and Hiremath and Salimath, 1992) although *E. tristachya*, a diploid (2n=2x=18) has not been eliminated as the maternal progenitor while *E. floccifolia*, a diploid (2n = 2x = 18) perennial species or an unknown or extinct ancestor is thought to be the paternal genome donor (Bisht and Mukai, 2000; Bisht and Mukai, 2001a and Bisht and Mukai, 2002; Liu et al., 2014). However, for these studies, the evidence was not enough since they only used one or several chloroplast genes or single low copy nuclear gene or marker. A minimal gene set or markers cannot give us convincing results. Thus, our objective was to attempt to study the heritage of *E. coracana* through *Eleusine* transcriptome sequencing and to construct a synthetic B transcriptome.

Materials and methods

Germplasm was acquired from the U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search.aspx) Germplasm Resources Information Network (NPGS GRIN) for analysis. An exhaustive search for all available *Eleusine* species was conducted to identify all possible candidate species within the *Eleusine* genus. Seven of the nine known *Eleusine* species were identified and acquired for analysis (Table 1). *Eleusine jaegeri* and *Eleusine kigeziensis* were unavailable from NPGS GRIN. No other sources for these two species could be identified. A previously assembled transcriptome, plastid genome, and mitochondrial genome of *E. indica* were utilized (Chen et al., 2015a; Zhang et al., 2017).

Eleusine species were germinated and grown from seed in a glasshouse environment at 28±2°C, and 70% average relative humidity in Auburn, AL (32.35°N, 85.29°W). Seedlings were grown in a native Wickham sandy loam soil with pH 6.3 and 0.5% organic matter. Four-week old entire seedlings were used for RNA extraction. Total RNA was extracted from individual seedlings of *E. multiflora*, *E. floccifolia*, *E. tristachya*, *E. intermedia*, *E. africana* and *E. coracana* using RNeasy Plant Mini Kit

(Qiagen, CA, USA). The quality and quantity of total RNA were determined with gel electrophoresis and Nanodrop 2000 (Thermo Scientific). High-quality RNA was used for transcriptome sequencing.

RNA preparation and sequencing was conducted at the Genomic Service Laboratory at Hudson Alpha Institute for Biotechnology (Cummings Research Park, Huntsville, AL) using standard procedures for the Illumina HiSeq 2000 to produce 100 bp paired-end reads (Chen et al., 2015a and Chen et al., 2015b). One complementary DNA (cDNA) library was constructed for each of the six total RNA samples. All samples were subjected to polyA selection prior to sequencing. *E. indica* transcriptome (NCBI Accession No.: SRR1560465) previously assembled by our lab (Chen et al., 2015a) was also sequenced by Hudson Alpha using the Illumina HiSeq 2000 platform and similar methodology. Besides, the samples were from same tissues (four-week old entire seedlings) with *E. indica* and in same growth conditions.

Sequence data analysis and assembly. Raw reads quality were checked by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then processed by Trimmomatic v.0.33 (Bolger et al., 2014; http://www.usadellab.org/cms/?page=trimmomatic) to remove adaptors and low qualified reads and sequences. The trimmed reads were evaluated with FastQC again and normalized with Trinity's in silico read normalization (http://trinityrnaseq.github.io), with maximum coverage of 30. Three *de novo* transcriptome assemblers were used: Trinity 2014-04-13p1 (Grabherr et al., 2011; http://trinityrnaseq.github.io), Velvet 1.2.08_ maxkmer101 (Zerbino and Birney, 2008; https://www.ebi.ac.uk/~zerbino/ velvet/), and SOAPdenovo2 v2.04 (Luo et al., 2012; https://github.com/aquaskyline/SOAPdenovo2). Trinity k-mer size was 25; Velvet k-mer size was 21 to 91 with step size of 10 and minimum contig length was 200 bp without scaffolding. SOAPdenovo2 k-mer size was 21 and 31. The three *de novo* assembler thus yielded 11 total assemblies for each species. The script Select_contigs.pl (https://pods.iplantcollaborative.org/wiki/display/DEapps/Select+contigs) was used for Trinity and SOAPdenovo2 to select contigs with minimum length 200 bp. To evaluate the quality of the assembly, N50s and contig length distributions of the assemblies were calculated with the script

Count_fasta.pl (http: //wiki.bioinformatics.ucdavis. edu/index.php/Count_fasta.pl). Before merging, "N"s were removed from the assemblies and contigs shorter than 200 bp were discarded.

All assemblies were combined into one merged assembly for each species individually. The merged assembly was processed by EvidentialGene tr2aacds pipeline (http: //arthropods. eugenes.org/EvidentialGene/about/EvidentialGene_ trassembly_pipe.html). The EvidentialGene pipeline takes as input the transcript FASTA file produced by any of the transcript assemblers and generates coding DNA sequences (CDSs) and amino acid sequences from each input contig then uses fastanrdb to quickly reduce perfect duplicate sequences, cd-hit and cd-hit-est to cluster protein and nucleotide sequences, and BLASTn and makeblastdb to find regions of local similarity between sequences. It outputs transcripts into three classes: *Okay* (the best transcripts with the unique CDS, which is close to a biologically real set regardless of how many millions of input assemblies.), *Alternate* (possible isoforms), and *Drop* (the transcripts did not pass the internal filter). The unique CDS (*Okay* set) and possible isoforms (*Alternate* set) were used for further evaluation and annotation. The overall workflow was summarized graphically in Figure 1.

Annotation and analysis. Sequences were annotated using Trinotate (Trinotate 2.02; https://trinotate.github.io/), which is a comprehensive annotation suite designed for automatic functional annotation of transcriptomes, particularly *de novo* assembled transcriptomes (Li et al., 2014). This pipeline includes: homology search to known sequence data (BLAST+/SwissProt), protein domain identification (HMMER/PFAM), protein signal peptide and transmembrane domain prediction (signalP/tmHMM), and leveraging various annotation databases (eggNOG/GO/Kegg databases). All functional annotation data derived from the analysis of transcripts are integrated into a SQLite database which allows fast efficient searching for terms with specific qualities related to a desired scientific hypothesis or a means to create a whole annotation report for a transcriptome.

Variants analysis. Variants are mainly classified into five different types: single nucleotide variants (SNVs), multiple nucleotide variants (MNVs), insertions, deletions, and replacements. SNVs are one base is replaced by another base, most commonly referred to as a single nucleotide polymorphism (SNP). MNVs are two or more SNVs in succession. Insertions are an event where one or more bases are inserted in the experimental data compared to the reference. Deletions are events where one or more bases are deleted from the experimental data compared to the reference. Replacements are more complex events where one or more bases have been replaced by one or more bases, where the identified allele has a length different from the reference.

Read mapping and detection of SNVs, MNVs, replacements, insertions, and deletions were conducted using the tools 'map reads to reference' and 'probabilistic variant detection' separately in CLC Genomics Workbench 6.5.2 (CLC Bio, Aarhus, Denmark, http://www.clcbio.com). The mapping parameters were set to 'Mismatch cost=3, Insertion cost=3, Deletion cost=3, Length fraction=0.95, Similarity fraction=0.95'. The variants calling parameters were set to 'Minimum coverage=30, Variant probability=90'.

Chloroplast and mitochondrial gene comparison. Complete *E. indica* chloroplast genome (KU833246) and mitochondrial genome (MF616338) were downloaded from NCBI. The *Eleusine* species' CDS datasets were aligned to the two genomes using Blastn at the E-value threshold 10^{-5}, word size 20, and minimum match size 90. *Eleusine coracana* reads were mapped to the aligned *Eleusine* species' CDSs separately. To verify the results, we also extracted the accurate CDS from *E. indica* chloroplast genome (KU833246) and mitochondrial genome separately. Chloroplast and mitochondrial genes of *E. indica* (NCBI) means these genes downloaded from NCBI, which were accurate assembled and uploaded before. However, genes of *E. indica* (transcriptome) were got using same blast method with other species and we can also use this method to verify our result. SNVs, MNVs, replacements, insertions, and deletions were called from each of the mappings in CLC Genomics Workbench 6.5.2

(CLC Bio, Aarhus, Denmark, http://www.clcbio.com). The SNVs, MNVs, replacements, insertions, and deletions were normalized by the total mapping consensus length.

Phylogenetic analysis. tBLASTx was used to extract chloroplast genes, mitochondrial genes from each *Eleusine* species separately. Genes were concatenated by FASconCAT-G_v1.02.pl (Kück and Meusemann, 2010) and one Supermatrix file was output which concatenated different kinds of species and genes. Ortholog pipelines were used for nuclear tree construction: the contigs were translated to coding protein sequences using Transdecoder v 3.0.1 (Ravin et al., 2016) following identification of the longest ORFs; Python script reduce_protein_redundancy.py (https://github.com/mcelrjo/blastp_nr) was used to get unique proteins. Orthofinder v1.1.8 (Emms and Kelly, 2015) was used to find orthogroups; A codon by gene partition scheme was used in Partition-Finder v2.0.0 (Lanfear et al., 2017) and model selection was limited to GTR-GAMMA and GTR-GAMMA+I with greedy search algorithm, and the best scheme was used for subsequent phylogenetic analysis. Trees were created using RAxML-MPI-AVX v8.2.6 (Stamatakis, 2014) with 100 rapid bootstraps, and GTRGAMMA model. Trees were visualized with Figtree (Rambaut, 2009).

Two diploid genome donors. *E. indica* chloroplast and mitochondrial genes were removed from the *E. coracana* reads. Then, the filtered *E. coracana* reads were mapped to *E. indica* transcriptome (unique CDS) using bowtie2 (Langmead and Salzberg, 2012) and then the mapped and unmapped reads were extracted using samtools (Li et al., 2009) separately. The mapped and unmapped reads were filtered and reads length below 90 were discarded. After filtering, the reads were divided into three parts: single reads, pair-end identical reads (forward read length is equal to reverse read length) and pair-end variant reads (forward and reverse read lengths are not equal). Both mapped and unmapped reads were assembled separately using Trinity, Velvet and SOAPdenovo2 and then subjected to the EvidentialGene tr2aacds pipeline. Then, *E. coracana* reads mapped to *E. indica* are referred to as the *E. coracana* Synthetic A

transcriptome and unmapped *E. coracana* are referred to as the *E. coracana* Synthetic B transcriptome. The other five *Eleusine* species' transcriptome reads were mapped to the *E. coracana* Synthetic B transcriptome and mapped percentage were calculated.

Results and discussion

Transcriptome sequencing and de novo assemblies. Read counts before and after quality checking and trimming are presented in Table 2. The summary statistics of the assemblies from EvidentialGene tr2aacds pipeline are shown in Table 3. Previous research has demonstrated this pipeline to improve transcript integrity and reduce assembly redundancy in transcriptome assembly (Chen et al., 2015a). Average read length after trimming was 99.31 to 99.42 nucleotides. The N50 of the unique CDS set ranged from 1471 to 1693, however, when the possible isoform set is added, the N50 ranged from 1232 and 1451.

For annotation, unique CDS assemblies of each transcriptome set were initially assigned with Trinotate. GoTermParse.py was used to retrieve GO Terms and three components (Table 4). GoTermParse.py (https://gist.github.com/NDHall/da9f9b9b3825bac7f1cb7508d4fec86e) used regular expressions and a dictionary to sort terms into their major functional groups (see supplementary materials). The GO classification assigned totals of 516,793; 634,349; 578,631; 803,545; 996,369; 1,039,581; 276,976; 243,115 and 697,893 GO terms to *E. multiflora, E. floccifolia, E. tristachya, E. intermedia, E. Africana, E. coracana, E. indica, E. coracana* Synthetic A transcriptome, and *E. coracana* Synthetic B transcriptome unique CDS set, respectively. The top ten GO terms of each component were compared among all species (Figure 2). All the GO terms in *E. coracana* 'unique CDS' set have higher scores than in others. Integral_component_of_membrane, transcription_DNA-templated and ATP_binding are the highest GO terms in each corresponding component.

The sequencing reads of *E. multiflora*, *E. floccifolia*, *E. tristachya, E. intermedia*, *E. africana, E. coracana* and *E. coracana* synthetic B were deposited at NCBI Sequence Read Archive (SRA) database

under the accessions SRR5467257, SRR5468569, SRR5468570, SRR5468571, SRR5468572, SRR5468573 and SRR6984602, respectively. Transcriptome Shotgun Assembly projects have been deposited at DDBJ/EMBL/GenBank under the accessions GGLR00000000, GGME00000000, GGMD00000000, GGMC00000000, GGMB00000000, GGMA00000000, and GGLY00000000, correspondingly. All of the versions described here are the first version, GGLR01000000, GGME01000000, GGMD01000000, GGMC01000000, GGMB01000000, GGMA01000000, and GGLY01000000.

Maternal genome donor. In order to elucidate the maternal genome donor of *E. coracana*, *E. coracana* reads were mapped to assembled and identified chloroplast and mitochondrial genes of *E. multiflora*, *E. floccifolia, E. tristachya, E. intermedia, E. africana*, *E. coracana*, *E. indica* (transcriptome) *and E. indica* (NCBI), respectively. *E. coracana* reads were also mapped to its own assembled and identified chloroplast and mitochondrial genes (Table 5). A total of 749,214; 943,716; 1,028,156; 1,342,048; 794,951; 1,354,667; 2,607,424; and 359,990 reads were mapped to chloroplast genes of *E. multiflora, E. floccifolia, E. tristachya, E. intermedia, E. africana*, *E. coracana*, *E. indica* (transcriptome) *and E. indica* (NCBI), respectively, and covered 76,737; 104,665; 94,699; 91,367; 125,454; 120,476; 89,464; and 59,475 bp of the references, respectively. Totals of 698,238; 1,943,279; 2,292,287; 2,279,808; 3,917,637; 1,199,393; 2,470,000; and 181,322 reads were mapped to mitochondrial genes and covered 99,312; 147,014; 136,351; 193,146; 210,108; 204,021; 87,813; and 38,146 bp of the references correspondingly (Table 5). The length of variants (SNVs, MNVs, replacements, insertions, and deletions) per million base pairs consensus detected from the *E. coracana* reads mapping to the chloroplast and mitochondrial genes of *Eleusine* species were calculated. The least total variants were mappings of *E. coracana* reads to *E. indica* chloroplast genes, using both *E. indica* (transcriptome) and *E. indica* (NCBI). Variants from both *E. indica* data sets were lower than mapping *E. coracana* reads to its own chloroplast genes. Similar results were found in mitochondrial genes. *E. coracana* reads mapping to *E. indica* mitochondrial genes

showed the least variants per million base pair consensus length than *E. coracana* reads to other *Eleusine* species, even the polyploid species *E. africana*.

In addition, concatenated phylogenetic trees were rooted using chloroplast, mitochondrial and ortholog genes, separately (Figure 3A, 3B and 3C). For the chloroplast derived tree, *E. coracana*, *E. africana* and *E. indica* (both *E .indica*_transcriptome and chloroplast sequences) were in the same branch with *Cyperus esculentus*, *Oryza sativa*, *Brachypodium sylvaticum*, *Paspalum urvillei* and *Zea mays* as outgroups, which strongly support that *E .indica* as maternal genome donor. Mitochondrial and nuclear tree analysis does rule out *E. floccifolia*, *E. intermedia*, and *E. multiflora* as potential maternal genome donors, and we have high bootstrap to support for this. It does not rule out *E. indica* or *E. tristachya* as the maternal genome donor. But given the slow rate of evolution within mitochondrial and nuclear sequences, this is not surprising. As such, these results indicated that *E. indica* is the maternal genome donor of *E. coracana*. Our maternal genome donor conclusion is consistent with most of other researches using different methods such as Genomic *in situ* hybridization (GISH), cytogenetic analysis and phylogenetic analysis to conclude *E. indica* is the maternal parent of *E. coracana* (Bisht and Mukai, 2001a, b). Besides, Hatakeyama et al. (2017) also constructed a molecular phylogenetic analysis using the detected low-copy-number homeologs during *E. coracana* genome assembly, and *E. indica* was close to *E. coracana*, which is consistent with our phylogenetic analysis. Draft *E. coracana* genome sequence is a good genomic resource for further genomic research of this species at the molecular level (Hittalmani et al., 2017; Hatakeyama et al., 2017). Compared to other traditional methods, using chloroplast genome is a more convenient, less time consuming, and a reliable tool for inferring phylogenetic relationships in polyploid species (Hilu, 1988). Chloroplast DNA is highly conserved and its potential usefulness in phylogenetic studies has been well documented (Curtis and Clegg, 1984; Palmer, 1985; Hilu, 1988). Considering the morphological characters, in addition *E. indica* and wild finger millet (subsp. africana) are highly similar (Hilu and De Wet, 1976). Here, we broadened the *E. coracana* maternity analysis to all assembled chloroplast and mitochondrial genes in our all *Eleusine* transcriptome profiles. In addition, a sister

relationship between *E. multiflora* and *E. floccifolia* was strongly supported by all of the phylogenetic trees.

Paternal genome donor. To decipher the paternal parent of *E. coracana*, our methodology was aimed at creating two Synthetic progenitor genomes by using the maternal parent of *E. indica* as a filter (Figure 4). First, using *E. indica* chloroplast and mitochondrial assembled transcripts, *E. coracana* reads were filtered to remove chloroplast and mitochondrial reads. The filtered *E. coracana* reads were then mapped to the *E. indica* transcriptome (unique CDS) using bowtie2 and the unmapped reads were extracted using samtools. The mapped percentage was 67.3% and properly paired was 62.6%, which also indicated that *E. indica* is the maternal genome donor compared to *E. coracana* read mapping to other transcriptomes. By filtering reads two distinct read sets were generated to create *E. coracana* synthetic A and synthetic B transcriptomes. The synthetic A transcriptome represents the maternal genome donor of *E. indica* and the synthetic B transcriptome represents the paternal genome donor of unknown origin. The unmapped reads were filtered and reads length below 90 were discarded. The mapped and unmapped pair-end reads were assembled by Trinity, Velvet and SOAPdenovo2 and then subjected to the EvidentialGene tr2aacds pipeline as described previously. The other five *Eleusine* species' reads were mapped to the Synthetic B transcriptome and mapped percentage were calculated (Table 6). By comparing mapping percentage, *E. floccifolia* mapped only 42.96% possibly indicating it is not the paternal genome donor of *E. coracana*. Filtered *E. coracana* reads mapped to *E. indica* were used same pipeline to get Synthetic A transcriptome. *E. indica*'s reads were mapped to this Synthetic A transcriptome and the mapped percentage is 72.90%. If one of the five *Eleusine* species is the paternal genome donor, the mapped percentage should be similar to *E. indica*'s read mapping to the Synthetic A transcriptome. However, except for *E. africana*, the mapped percentages of the other four species are less than 50% (Table 6). Further, variants produced by mapping to Synthetic B transcriptome of all species were greater compared to *E. indica* mapping to Synthetic A transcriptome (Table 6).

The SNPs of *E. tristachya*'s were lower than *E. flocciflolia*'s, so the latter one is not a genome donor. If one species is a B genome donor in these five species, the SNPs of the species should obviously lower than other four species. However, the variants of *E. multiflora*, *E. floccifolia* and *E. tristachya* are not appearing to be much different.

Phylogenetic concatenated nuclear tree was obtained using all of the ortholog genes (Figure 3C). The Synthetic B transcriptome was grouped with *E. africana* with low bootstrap value (value = 56) which suggests there is no extant B genome donor. Since *E. coracana* was domesticated from wild species *E. africana*, which is also an allotetraploid species and it cannot be the diploid genome donor (Bisht and Mukai, 2002). The Synthetic A transcriptome grouped with *E. indica*, *E. tristachya* and *E. coracana*.

The purpose of this study is not only to construct the transcriptome references for *Eleusine* species, but also to construct the synthetic B transcriptome of *E. coracana*. *E. floccifolia* was denied as potential B genome donor. The synthetic B transcriptome will be able to aid future research. Combine the results from this project and *E. coracana* genome information will be useful for the continued study of *E. coracana*.

Estimate the divergence time of E. coracana. More than 9000 current species of grasses were derived from a common ancestor that lived about 50-80 million years ago (mya) (Prasad, 2011; Crepet and Feldman, 1991). People used paleontological, like phytolith to study the origin of grass, however, it has prevented detailed examination of their evolution as lack of an early fossil record of grasses (Huang et al. 2007). Therefore, we have very little direct knowledge about the timing of evolution of grasses. The domestication of crop grass began c. 12 000 yr ago in the Fertile Crescent (Glémin and Bataillon, 2009). Extreme variations in paleoclimates might explain the observed divergence in allopolyploid lineages (Stebbins, 1980). Liu et al. (2011) used six plastid markers to estimate the divergence time of *E. coracana*. Their results indicated that the crown age of *Eleusine* was determined to be 3.89 mya in the Miocene-early Pliocene interval and the divergence of *E. coracana* was estimated to have occurred 0.67

mya in the late Pleistocene. Studies indicated that finger millet originated in the East African Highlands and was subsequently introduced into India. *E. africana* is a wild finger millet and its morphology is very similar to *E. indica* (Hilu et al., 1976).

Conclusion

These results suggest that *E. indica* as the A genome donor and the B genome donor is not one we included in this study. Considering that *E. jaegeri* and *E. kigeziensis* cannot be the paternal genome donor of *E. coracana* since the chromosome number of *E. jaegeri* is 20 and *E. kigeziensis* is also a tetraploid species with chromosome number 2n = 38. And the unsampled *E. semisterilis* also seems an unlikely candidate for paternal parents because of its unusual morphology of laxly arranged spikelets (Liu et al. 2011). There are some other hypotheses to explain why the paternal parents remain unidentified, such as the paternal parents of *E. coracana* may be from outside *Eleusine* and thus they remain unidentified because of restricted sampling at the intergeneric level. Another one is that the genome has undergone a great change after allotetraploid speciation and is untraceable now (Liu et al. 2011). The possibility that the paternal genome donor of *E. coracana* is extinct is consistent with other hypotheses (Hiremath and Salimath 1992; Neves et al. 2005; Devarumath et al. 2010; Liu et al. 2011).

In this study, we constructed optimized transcriptome references for *E. multiflora*, *E. floccifolia*, *E. tristachya*, *E. intermedia*, *E. Africana*, *E. coracana* and *E. coracana* synthetic B using three de novo assemblers and a redundancy-reducing pipeline. By comparing the chloroplast and mitochondrial genes among *Eleusine* species, we demonstrated that *E. indica* as maternal genome donor. *E. coracana* reads filtered for only chloroplast and mitochondrial genes were mapped to *E. indica* transcriptome and the unmapped reads were extracted and assembled. Other five *Eleusine* species' transcriptome reads were mapped to the *E. coracana* Synthetic B transcriptome and mapped percentage and variants were compared, however, we found that the mapped percentage were very low and there were many variants between *E. coracana* Synthetic B transcriptome and each other *Eleusine* species. Besides, phylogenetic

analyses using ortholog genes also suggest that no *Eleusine* species close to the *E. coracana* Synthetic B transcriptome branch. Evidence suggests the B genome donor is extinct or is another unidentified species. Transcriptomes are made publically available for comparison to other species to aid in discovery of the B genome donor if it still exists. Abundant genetic resources and the *E. coracana* synthetic B transcriptome from this research will be useful for the continued study of *E. coracana* and plant breeding.

Table 1 Biological, genomic, and GRIN Accession Number for seven *Eleusine* species utilized. Genomic and biological acquired from the following sources: †

Species	2n chromosome numbers, Genome, ploidy	Life cycle	Type	GRIN Accession Number
E. multiflora	16, CC, diploid	Annual	Wild	226067
E. floccifolia	18, BB, diploid	Perennial	Wild	196853
E. tristachya	18 AA, diploid	Annual	Wild	331791
E. intermedia	18 AB, diploid	Perennial	Wild	273888
E. africana	36 AABB, allotetraploid	Annual	Wild	226270
E. coracana	36 AABB, allotetraploid	Annual	Cultivated	462949
E. indica	18 AA, diploid	Annual	Wild	Collect ‡

† GRIN, Germplasm Resources Information Network

‡ *E. indica* was collected locally from a crop field in Tallassee, Alabama. In other published work by J.S. McElroy it is referred to by the acronym PBU referring to its origin at the Alabama Agricultural Experiment Station Plant Breeding Unit.

Table 2 The number and average length of *Eleusine* transcriptome sequencing reads before and after trimming

Species	Number of reads	Average length	Number of reads after trim	% trimmed reads	Average length after trim
E. multiflora	61,348,758	100	52,236,532	15%	99.41
E. floccifolia	59,140,884	100	50,053,954	15%	99.40
E. tristachya	53,661,434	100	45,004,810	16%	99.42
E. intermedia	106,867,304	100	84,798,308	21%	99.40
E. africana	197,003,984	100	156,392,016	21%	99.34
E. coracana	139,928,698	100	111,917,028	20%	99.31
E. indica	230,466,942	100	183,323,866	17%	99.39

Table 3 N50, sequences number and total length of the assemblies in EvidentialGene tr2aacds pipeline

Species	Unique CDSs			Unique CDSs + Possible isoforms		
	N50 (bp)	Sequences number	Total length (bp)	N50 (bp)	Sequence number	Total length (bp)
E. multiflora	1567	30,394	32,083,609	1357	52,610	50,466,628
E. floccifolia	1585	36,364	37,932,847	1361	72,602	69,442,718
E. tristachya	1549	35,856	37,243,265	1353	72,764	69,722,866
E. intermedia	1693	39,540	43,739,409	1451	87,270	87,954,199
E. africana	1516	56,375	54,910,276	1236	144,921	129,354,728
E. coracana	1471	59,223	561,062,47	1232	144,460	128,133,958
E. indica	1562	25,878	28,239,951	1408	36,959	37,055,659

Table 4 The number and percentage of total of cellular component, molecular function, biological process of GO terms in nine transcriptome data sets. Each contig has multiple GO terms during annotation leading to the number of GO terms more than contigs

Species	Cellular component	Molecular function	Biological process	Total GO
E. multiflora	142,392 (27.6%)	180,998 (35.0%)	193,403 (37.4%)	516,793
E. floccifolia	172,785 (27.2%)	224,107 (35.3%)	237,457 (37.4%)	634,349
E. tristachya	155,857 (26.9%)	205,303 (35.5%)	217,471 (37.6%)	578,631
E. intermedia	216,357 (26.9%)	281,597 (35.0%)	305,591 (38.0%)	803,545
E. africana	269,158 (27.0%)	355,928 (35.7%)	371,283 (37.3%)	996,369
E. coracana	275,677 (26.5%)	374,256 (36.0%)	389,648 (37.5%)	1,039,581
E. indica	76,620 (27.7%)	94,054 (34.0%)	106,302 (38.4%)	276,976
Synthetic A	66,621 (27.4%)	84,636 (34.8%)	91,858 (37.8%)	243,115
Synthetic B	193,781 (27.8%)	241,307 (34.6%)	262,805 (37.7%)	697,893

Table 5 The mapped reads, covered references, mapped percentage and the length of SNVs, MNVs, replacements, insertions, and deletions per million base pairs consensus detected from the *E. coracana* reads mapped to the chloroplast and mitochondrial genes of other *Eleusine* species

	Chloroplast							
Species	Mapped reads	Covered references	Mapped percentage	SNVs	MNVs	Replacements	Insertions	Deletions
E. coracana	1,354,667	120,476	1.14%	1289	33	0	25	42
E. multiflora	749,214	76,737	0.63%	2501	82	0	190	190
E. floccifolia	943,716	104,665	0.80%	2624	99	20	129	257
E. tristachya	1,028,156	94,699	0.87%	1478	43	21	54	278
E. intermedia	1,342,048	91,367	1.13%	5599	47	82	305	200
E. africana	794,951	125,454	0.67%	1848	65	0	114	131
E. indica (transcriptome)	2,607,424	89,464	2.20%	819	23	0	0	114
E. indica (KU833246)	359,990	59,475	0.30%	354	0	0	0	0
	Mitochondria							
Species	Mapped reads	Covered references	Mapped percentage	SNVs	MNVs	Replacements	Insertions	Deletions
E. coracana	1,199,393	204,021	1.01%	1560	40	0	30	25
E. multiflora	698,238	99,312	0.59%	4468	125	45	159	239
E. floccifolia	1,943,279	147,014	1.64%	3604	125	0	86	133
E. tristachya	2,292,287	136,351	1.94%	2614	92	38	69	77
E. intermedia	2,279,808	193,146	1.92%	6246	135	78	284	156
E. africana	3,917,637	210,108	3.31%	2426	151	0	111	75
E. indica (transcriptome)	2,470,000	87,813	2.09%	1949	71	0	106	47
E. indica (MF616338)	181,322	38,146	0.15%	1530	0	0	0	0

Table 6 The mapped percentage, mapped reads, covered references and the length of SNVs, MNVs, replacements, insertions, and deletions per million base pairs consensus detected from the *Eleusine* species' reads mapped to the *E. coracana* Synthetic B transcriptome

Species	Mapped reads	Covered† references	Mapped percentage	SNVs	MNVs	Replacements	Insertions	Deletions
E. coracana	64,503,333	38,974,337	54.45%	2958	86	7	83	88
E. multiflora	23,104,445	38,974,337	41.83%	1168	27	3	31	35
E. floccifolia	22,722,647	38,974,337	42.96%	1285	30	4	36	43
E. tristachya	22,526,872	38,974,337	47.16%	979	22	4	29	39
E. intermedia	14,857,792	38,974,337	16.28%	2975	63	4	39	42
E. africana	87,841,046	38,974,337	52.78%	2833	71	7	83	87

†The covered references are same because all of the *Eleusine* species' reads were mapped to the *E. coracana* Synthetic B transcriptome.

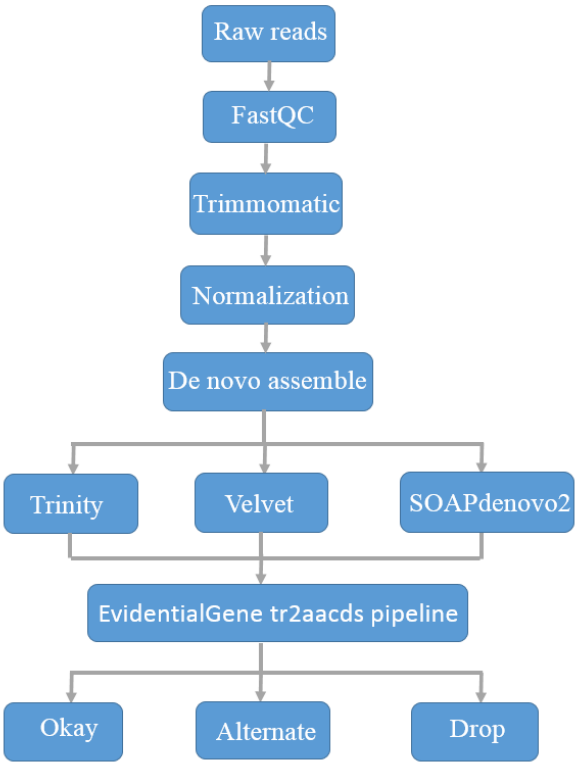


Figure 1 Workflow of transcriptome sequencing data analysis and assembly. Three de novo assemblers (Trinity, Velvet, and SOAPdenovo2) and a redundancy-reducing EvidentialGene tr2aacds pipeline were used for constructing optimized transcriptome references.

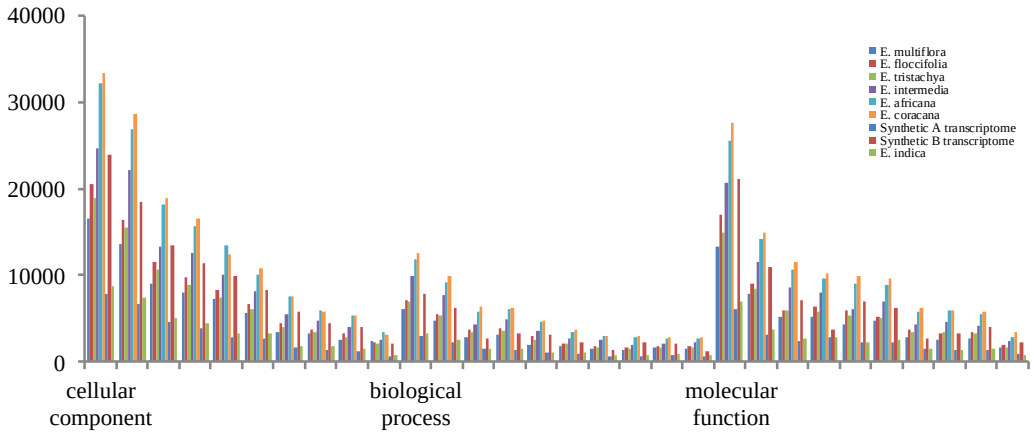
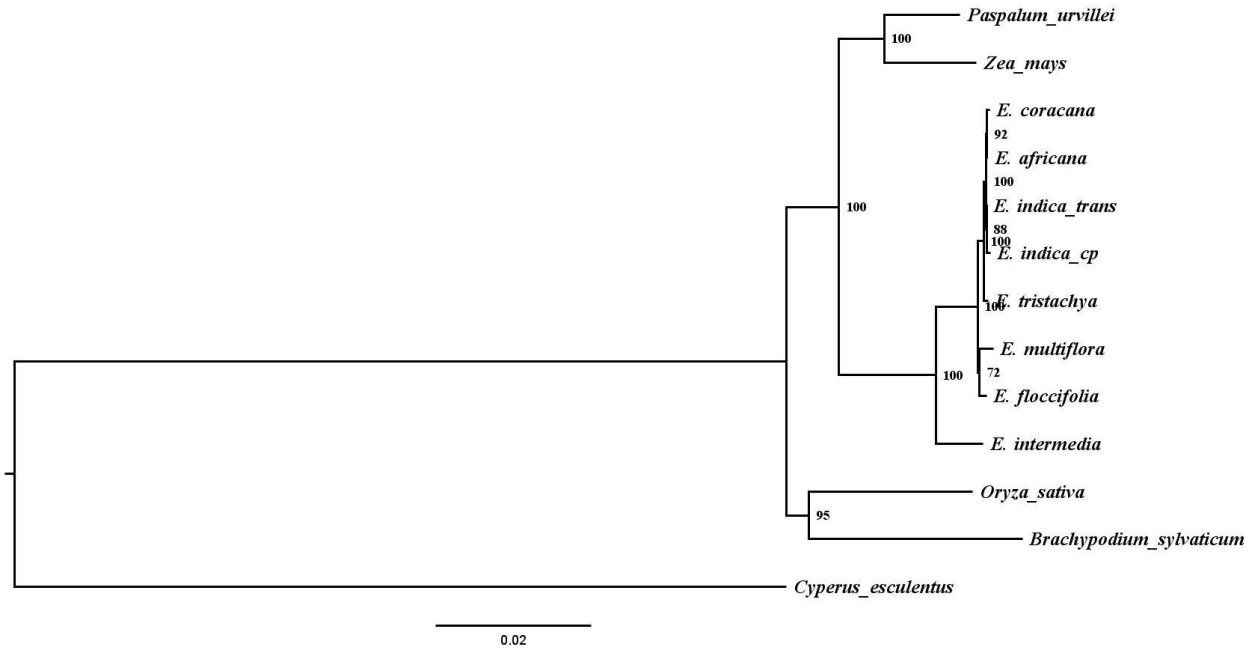
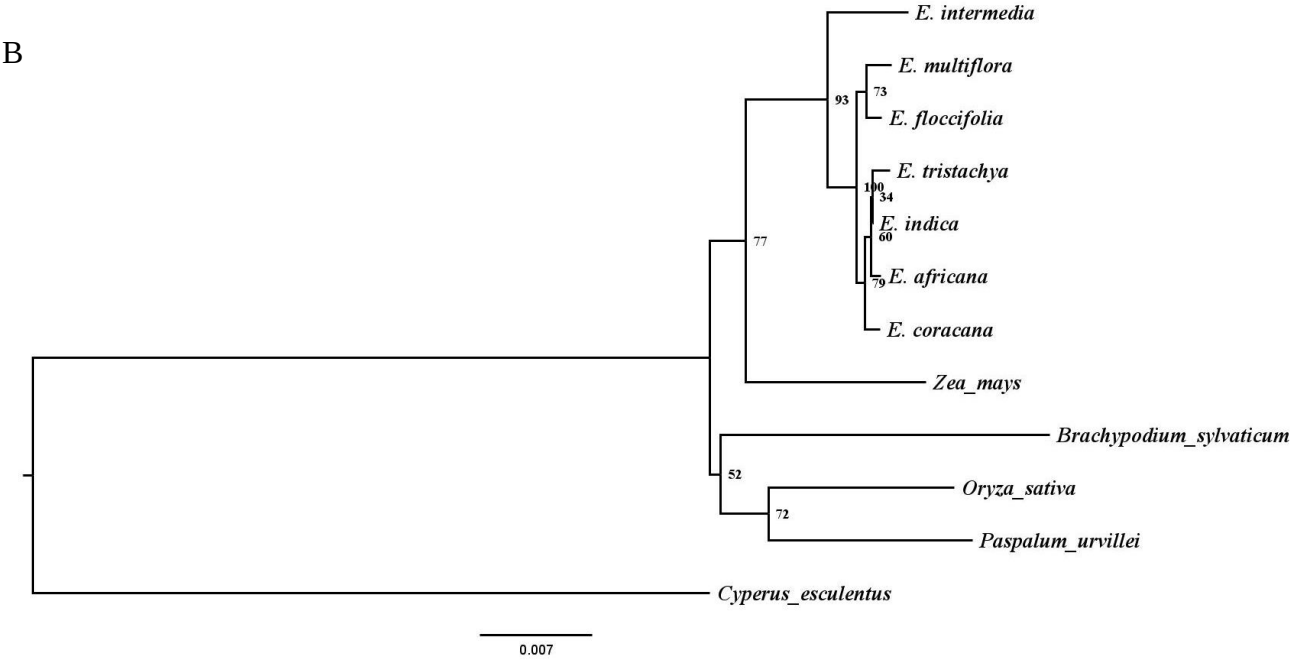


Figure 2 GO classifications of all *Eleusine* species, Synthetic A and B transcriptomes. The results were summarized in three main categories: biological process, cellular component and molecular function. The top ten GO terms of each component were compared among all *Eleusine* species, Synthetic A and B transcriptomes.

A


Paspalum_urvillei
100
Zea_mays
E. coracana
92
E. africana
100
E. indica_trans
88
100
E. indica_cp
100
E. tristachya
100
E. multiflora
72
E. floccifolia
100
E. intermedia
Oryza_sativa
95
Brachypodium_sylvaticum
Cyperus_esculentus
0.02

B


E. intermedia
E. multiflora
73
93
E. floccifolia
E. tristachya
100
34
E. indica
60
77
79
E. africana
E. coracana
Zea_mays
Brachypodium_sylvaticum
52
Oryza_sativa
72
Paspalum_urvillei
Cyperus_esculentus
0.007

C

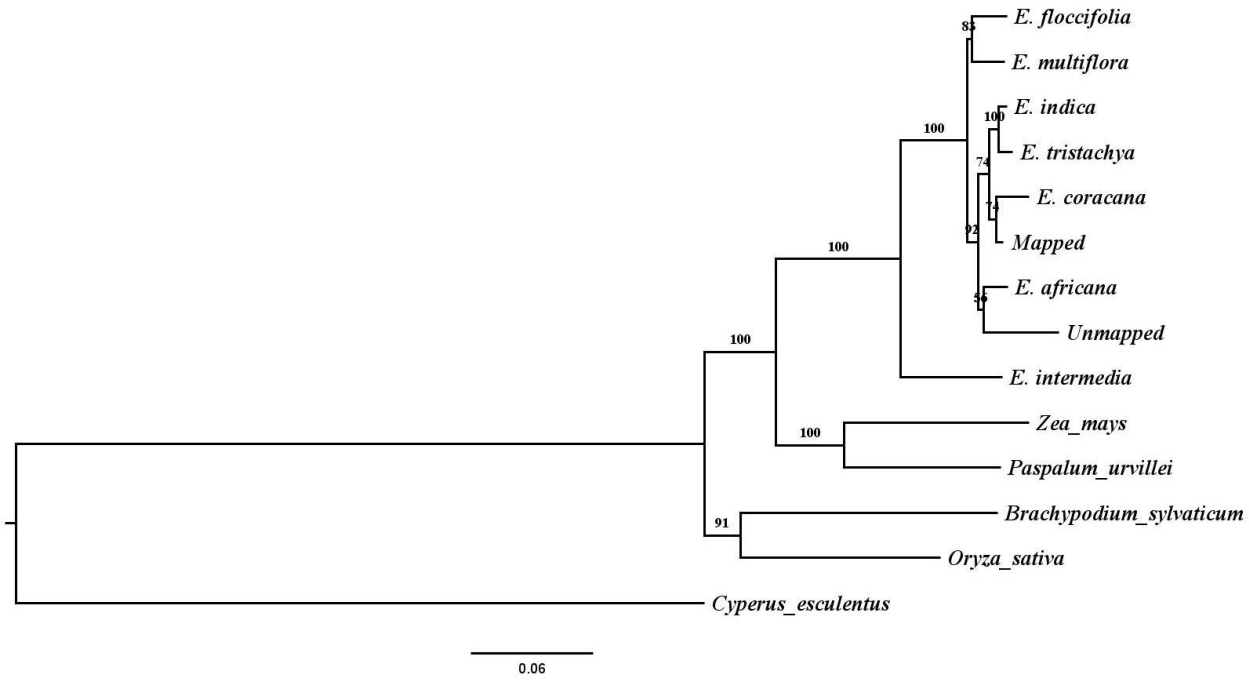


Figure 3 (A) Phylogenetic trees made using concatenated chloroplast genes in RAxML. (B) Phylogenetic trees made using concatenated mitochondrial genes in RAxML. (C) Tree based on orthologous genes. Mapped: Synthetic A transcriptome; Unmapped: Synthetic B transcriptome.

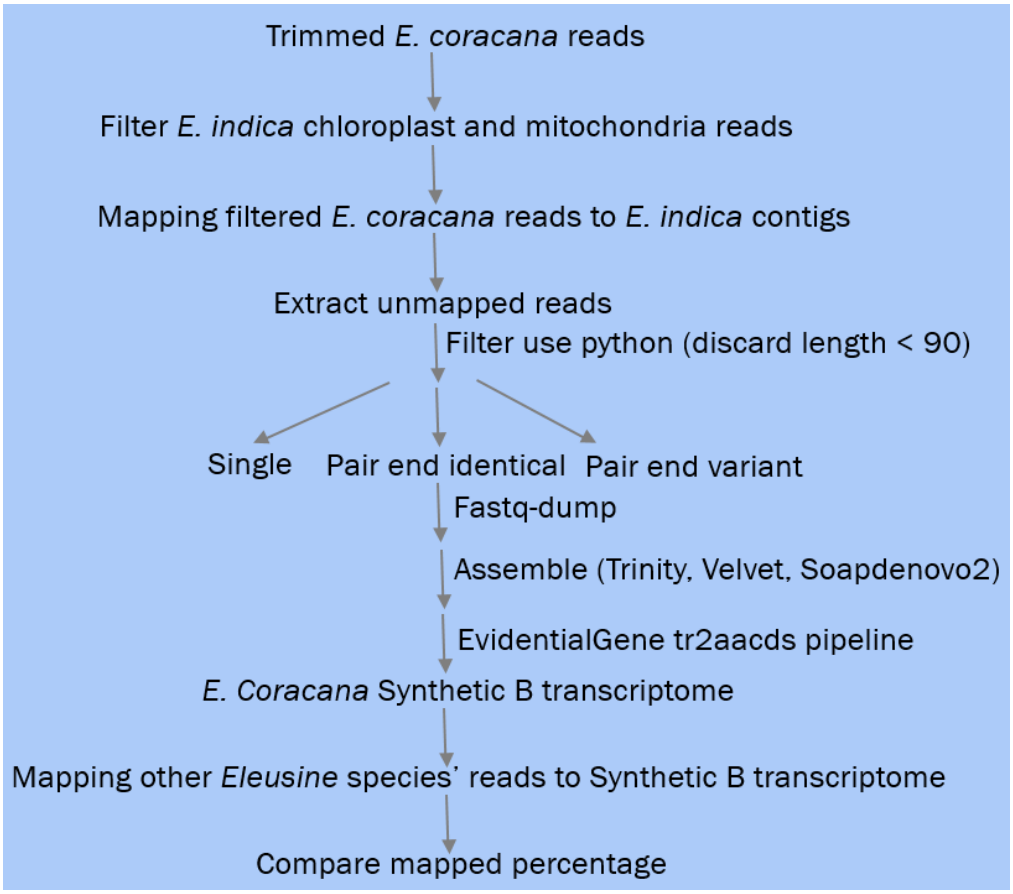


Figure 4 The pipeline of determining B genome donor of *E. coracana*. Filtered *E. coracana* reads were mapped to *E.indica* transcriptome and extracted the unmapped reads and assembled. Other five *Eleusine* species' transcriptome were mapped to the *E. coracana* Synthetic B transcriptome and mapped percentages were calculated.

References

Acland, A., R. Agarwala, T. Barrett, J. Beck, D.A. Benson, C. Bollin, E. Bolton, S.H. Bryant, K. Canese, D.M. Church, K. Clark, M. Dicuccio, I. Dondoshansky, S. Federhen, M. Feolo, L.Y. Geer, V. Gorelenkov, M. Hoeppner, M. Johnson, C. Kelly, V. Khotomlianski, A. Kimchi, M. Kimelman, P. Kitts, S. Krasnov, A. Kuznetsov, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, I. Karsch-Mizrachi, T. Murphy, J. Ostell, C. O'Sullivan, A. Panchenko, L. Phan, D.P.K.D. Pruitt, W. Rubinstein, E.W. Sayers, V. Schneider, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, K. Siyan, D. Slotta, A. Soboleva, V. Soussov, G. Starchenko, T.A. Tatusova, B.W. Trawick, D. Vakatov, Y. Wang, M. Ward, W. John Wilbur, E. Yaschenko, and K. Zbicz. 2014. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 42(D1): 7–17.

Agrawal, R., N. Agrawal, R. Tandon, and S. N. Raina. 2014. Chloroplast genes as genetic markers for inferring patterns of change, maternal ancestry and phylogenetic relationships among Eleusine species. AoB Plants, 6, plt056.

An, J., X. Shen, Q. Ma, C. Yang, S. Liu, and Y. Chen. 2014. Transcriptome profiling to discover putative genes associated with paraquat resistance in goosegrass (Eleusine indica L.). PLoS One 9(6): e99940.

Basu, C., M. D. Halfhill, T. C. Mueller, and C. N. Stewart Jr. 2004. Weed genomics: new tools to understand weed biology. Trends in plant science, 9(8), 391-398.

Bhatnagar, S. S. 1952. The Wealth of India. Vol.III. Council of Scientific and Industrial.

Bisht, M. S., and Y. Mukai. 2001a. Genomic in situ hybridization identifies genome donor of finger millet (Eleusine coracana). Theor. Appl. Genet. 102: 825–832.

Bisht, M. S., and Y. Mukai. 2001b. Identification of genome donors to the wild species of finger millet, Eleusine africana by genomic in situ hybridization. Breed. Sc. 51: 263–269.

Bisht, M.S., and Y. Mukai. 2000. Mapping of rDNA on the chromosomes of Eleusine species by fluorescence in situ hybridization. Genes Genet. Syst. 75: 343–348.

Bisht, M.S., and Y. Mukai. 2002. Genome organization and polyploid evolution in the genus Eleusine (Poaceae). Plant Syst. Evol. 233(3–4): 243–258.

Boetzer, M., C. V. Henkel, H.J. Jansen, D. Butler, and W. Pirovano. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27(4): 578–579.

Boisvert, S., F. Laviolette, and J. Corbeil. 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. 17(11): 1519–1533.

Bolger, A.M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114–2120.

Bull, L.N., C.R. Pabón-Peña, and N.B. Freimer. 1999. Compound microsatellite repeats: practical and theoretical features. Genome Res. 9(9): 830–838.

Chen, S., J. S. McElroy, F. Dane, and L. R. Goertzen. 2015b. Transcriptome Assembly and Comparison of an Allotetraploid Weed Species, Annual Bluegrass, with its Two Diploid Progenitor Species, Schrad and Kunth. Plant Genome 0(0): 0Available at https://dl.sciencesocieties.org/publications/tpg/abstracts/0/0/plantgenome2015.06.0050.

Chen, M., G. Zeng, Z. Tan, M. Jiang, J. Zhang, C. Zhang, L. Lu, Y. Lin, and J. Peng. 2011. Compound microsatellites in complete Escherichia coli genomes. FEBS Lett. 585(7): 1072–1076.

Chen, S., J.S. McElroy, F. Dane, and E. Peatman. 2015a. Optimizing Transcriptome Assemblies for Leaf and Seedling by Combining Multiple Assemblies from Three De Novo Assemblers. Plant Genome 0(0): 0.

Chennaveeraiah, M. S., and S. C. Hiremath. 1974a. Genome analysis of Eleusine coracana (L.) Gaertn. Euphytica 23: 489–495.

Cho, K.-S., B.-K. Yun, Y.-H. Yoon, S.-Y. Hong, M. Mekapogu, K.-H. Kim, and T.-J. Yang. 2015. Complete Chloroplast Genome Sequence of Tartary Buckwheat (Fagopyrum tataricum) and Comparative Analysis with Common Buckwheat (F.esculentum). PLoS One 10(5): e0125332.

Cronn, R., A. Liston, M. Parks, D.S. Gernandt, R. Shen, and T. Mockler. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36(19): e122.

Cui, L. 2006. ChloroplastDB: the Chloroplast Genome Database. Nucleic Acids Res. 34(90001): D692–D696.

Curci, P.L., D. De Paola, D. Danzi, G.G. Vendramin, and G. Sonnante. 2015. Complete Chloroplast Genome of the Multifunctional Crop Globe Artichoke and Comparison with Other Asteraceae. PLoS One 10(3): e0120589.

Curtis, S. E., and M. T. Clegg. 1984. Molecular evolution of chloroplast DNA sequences. Mol. Biol. Evol. 1: 291-301.

Darmency, H. 1994. Genetics of herbicide resistance in weeds and crops. Herbicide Resistance in Plants: Biology and Biochemistry, 263-298.

Dawson, I.K., P.E. Hedley, L. Guarino, and H. Jaenicke. 2009. Does biotechnology have a role in the promotion of underutilised crops? Food Policy 34(4): 319–328Available at http://dx.doi.org/10.1016/j.foodpol.2009.02.003.

Devarumath, R M., S. S. Sheelavanthmath, and S. C. Hiremath. 2010. Chromosome pairing analysis in interspecific hybrids among tetraploid species of Eleusine (Poaceae). Indian Journal of Genetics 70: 299 –303.

Ehara, M., Y. Inagaki, K. I. Watanabe, and T. Ohama. 2000. Phylogenetic analysis of diatom coxI genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. Current genetics, 37(1), 29-33.

Emms, D.M., and S. Kelly. 2015. OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol.: 1–14Available at http://dx.doi.org/10.1186/s13059-015-0721-2.

Frazer, K.A., L. Pachter, A. Poliakov, E.M. Rubin, and I. Dubchak. 2004. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 32(Web Server): W273–W279.

Givnish, T.J., M. Ames, J. R. McNeal, M. R. McKain, P. R. Steele, S. W. Graham, J. C. Pires, D. W. Stevenson, W. B. Zomlefer, B. G. Briggs, and M. R. Duvall. 2010. Assembling the Tree of the Monocotyledons: Plastome Sequence Phylogeny and Evolution of Poales1. Annals of the Missouri Botanical Garden, 97(4), pp.584-616.

Góngora-Castillo, E., and C. R. Buell. 2013. Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Natural product reports, 30(4), 490-500.

Gouy, M., S. Guindon, and O. Gascuel. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27(2): 221–4.

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Palma, B.W. Birren, C. Nusbaum, K. Lindblad-toh, N. Friedman, and A. Regev. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 29(7): 644-652.

Hammer, K., and J. Heller. 1998. Promoting the conservation and use of underutilized and neglected crops. Schriften Genet Resour, 8: 223-227.

Hammer, K., J. Heller, and J. Engels. 2001. Monographs on underutilized and neglected crops. Genet. Resour. Crop Evol. 48(1): 3–5.

Hatakeyama, M., S. Aluri, M.T. Balachadran, S.R. Sivarajan, L. Poveda, R. Shimizu-inatsugi, A. Patrignani, S. Gru, J. Baeten, K. Francoijs, K.N. Nataraja, Y.A.N. Reddy, S. Phadnis, R.L.

Ravikumar, R. Schlapbach, S.M. Sreeman, and K.K. Shimizu. 2017. Multiple hybrid de novo genome assembly of finger millet , an orphan allotetraploid crop. DNA Res. 0(September): 1–9.

Haughn, G.W., Smith, J.; Mazur, B., Somerville, C. 1988. Transformation with a mutant Arabidopsis acetolactate synthase gene renders tobacco resistant to sulfonylurea herbicides. Mol. Gen. Genet., 211, 266-271.

He, S., Y. Wang, S. Volis, D. Li, and T. Yi. 2012. Genetic diversity and population structure: Implications for conservation of wild soybean (glycine soja sieb. et zucc) based on nuclear and chloroplast microsatellite variation. Int. J. Mol. Sci. 13(10): 12608–12628.

Heap, I. 2018. The International Survey of Herbicide Resistant Weeds. Online. Internet. Available at www.weedscience.org.

Hilu, K. W., and J. M. J. De Wet. 1976. Domestication of Eleusine coracana. Economic Botany, 30(3): 199-208.

Hilu, K. W. 1981. Taxonomic status of the disputable Eleusine compressa (Gramineae). Kew. Bull. 36: 559–562.

Hilu, K. W. 1988. Identification of the ‘‘A’’ genome of finger millet using chloroplast DNA. Genetics 118: 163–167.

Hilu, K. W., J. M. J. deWet, and D. R. Seigler. 1978. Flavonoid patterns and systematics in Eleusine. Biochem. Syst. Ecol. 6: 247-249.

Hiremath, S. C., and S. S. Salimath. 1991a. The quantitative nuclear DNA changes in Eleusine (Gramineae). Plant Syst. Evol. 178: 225–233.

Hiremath, S. C., and S. S. Salimath. 1992. The ‘‘A’’ genome donor of Eleusine coracana (L.) Gaertn. (Gramineae). Theor. Appl. Genet. 84: 747–754.

Hittalmani, S., H.B. Mahesh, M.D. Shirke, H. Biradar, G. Uday, Y.R. Aruna, H.C. Lohithaswa, and A. Mohanrao. 2017. Genome and Transcriptome sequence of Finger millet (Eleusine coracana (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. BMC Genomics 18(1): 465Available at http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3850-z.

Holm, L. G., D. L. Plucknett, J. V. Pancho, and J. P. Herberger. 1977. The world's worst weeds. University Press.

Horvath, D. 2010. Genomics for weed science. Current genomics, 11(1), 47-51.

Jansen, R.K., L.A. Raubeson, J.L. Boore, C.W. DePamphilis, T.W. Chumley, R.C. Haberle, S.K. Wyman, A.J. Alverson, R. Peery, S.J. Herman, H.M. Fourcade, J. V. Kuehl, J.R. McNeal, J.

Leebens-Mack, and L. Cui. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. 395: 348–384.

Kofler, R., C. Schlötterer, E. Luschützky, and T. Lelley. 2008. Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genomics 9: 612.

Kück, P., and K. Meusemann. 2010. FASconCAT: Convenient handling of data matrices. Mol. Phylogenet. Evol. 56(3): 1115–1118.

Kumar, A., V.S. Gaur, A. Goel, and A.K. Gupta. 2014. De Novo Assembly and Characterization of Developing Spikes Transcriptome of Finger Millet (Eleusine coracana): a Minor Crop Having Nutraceutical Properties. Plant Mol. Biol. Report. 33(4): 905–922.

Lanfear, R., B. Calcott, S.Y.W. Ho, and S. Guindon. 2017. PartitionFinder : Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses Research article. 29(September): 1695–1701.

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4): 357–359.

Lee, R.M., J. Thimmapuram, K.A. Thinglum, G. Gong, A.G. Hernandez, C.L. Wright, R.W. Kim, M.A. Mikel, and P.J. Tranel. 2009. Sampling the Waterhemp (Amaranthus tuberculatus) Genome Using Pyrosequencing Technology. 57:463-469. 57(5): 463–469Available at http://iwakami00462.pdf.

Li, C., B.H. Beck, S.A. Fuller, and E. Peatman. 2014. Transcriptome annotation and marker discovery in white bass (Morone chrysops) and striped bass (Morone saxatilis). : 885–887.

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(16): 2078–2079.

Liu, Q., B. Jiang, J. Wen, and P. M. Peterson. 2014. Low-copy nuclear gene and McGISH resolves polyploid history of Eleusine coracana and morphological character evolution in Eleusine. Turkish Journal of Botany, 38(1), 1-12.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. 2012. Comparison of next-generation sequencing systems. BioMed Research International, 2012.

Liu, Q., J.K. Triplett, J. Wen, and P.M. Peterson. 2011. Allotetraploid origin and divergence in Eleusine (Chloridoideae, Poaceae): Evidence from low-copy nuclear gene phylogenies and a plastid gene chronogram. Ann. Bot. 108(7): 1287–1298.

Lohse, M., O. Drechsel, and R. Bock. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. : 267–274.

Lowe, T.M., and S.R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5): 955–64.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D.W. Cheung, S. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. Lam, and J. Wang. 2012. SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler. : 1–6.

Jasieniuk, M., A.L. Brûlé-Babel, and I.N. Morrison. 1996. The evolution and genetics of herbicide resistance in weeds, Weed Sci. 44 176–193.

Ma. X., H. Wu, W. Jiang, and Y. Ma. 2015. Goosegrass (Eleusine indica) density effects on cotton (Gossypium hirsutum). Integr. Agric. 14(9): 1778–1785.

Mahesh, H.B., M.D. Shirke, S. Singh, A. Rajamani, S. Hittalmani, G. Wang, and M. Gowda. 2016. Indica rice genome assembly , annotation and mining of blast disease resistance genes. : 1–12.

Maier, R.M., and C. Schmitz-Linneweber. 2004. Plastid Genomes. In: Daniell H., Chase C. (eds) Molecular Biology and Biotechnology of Plant Organelles. Springer, Dordrecht

Maier, R.M., K. Neckermann, G.L. Igloi, and H. Kössel. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J. Mol. Biol. 251(5): 614–628.

Martin, G., F.C. Baurens, C. Cardi, A. D'Hont, and J.M. Aury. 2013. The Complete Chloroplast Genome of Banana (Musa acuminata, Zingiberales): Insight into Plastid Monocotyledon Evolution. PLoS One 8(6).

McCullough, P.E., J. Yu, J.S. McElroy, S. Chen, H. Zhang, T.L. Grey, and M.A. Czarnota. 2016b. ALS-Resistant Annual Sedge (Cyperus compressus) Confirmed in Turfgrass. Weed Sci. 64(1): 33–41.

McCullough, P.E., J.S. McElroy, J. Yu, H. Zhang, T.B. Miller, S. Chen, C.R. Johnston, and M.A. Czarnota. 2016a. ALS-Resistant Spotted Spurge (Chamaesyce maculata) Confirmed in Georgia. Weed Sci. 64(2): 216–222Available at https://www.cambridge.org/core/product/identifier/S004317450001540X/type/journal_article.

McElroy, J.S. 2016. Goosegrass seed biology Goosegrass environmental tolerances. : 1–5.

Mehra, K. L. 1963a. Differentiation of cultivated and wild Eleusine species. Phyton. 20: 189–198.

Mehra, K.L. 1962. Natural hybridization between Eleusine coracana and E. africana in Uganda. J. Indian Bot. Soc. 41: 531–539.

Melotto-Passarin, D.M., E. V. Tambarussi, K. Dressano, V.F. de Martin, and H. Carrer. 2011. Characterization of chloroplast DNA microsatellites from Saccharum spp and related species. Genet. Mol. Res. 10(3): 2024–2033.

Meyer, A. and A. C. Wilson. 1990. Origin of tetrapods inferred from their mitochondrial DNA affiliation to lungfish. J. Mol. Evol. 31:359-364.

Millen, R.S., R.G. Olmstead, K.L. Adams, J.D. Palmer, N.T. Lao, L. Heggie, T.A. Kavanagh, J.M. Hibberd, J.C. Gray, C.W. Morden, P.J. Calie, L.S. Jermiin, and K.H. Wolfe. 2001. Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. Plant Cell 13(3): 645–58.

Milne, I., G. Stephen, M. Bayer, P.J.A. Cock, L. Pritchard, L. Cardle, P.D. Shawand, and D. Marshall. 2013. Using tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. 14(2): 193–202.

Naylor, R.L., W.P. Falcon, R.M. Goodman, M.M. Jahn, T. Sengooba, H. Tefera, and R.J. Nelson. 2004. Biotechnology in the developing world: A case for increased investments in orphan crops. Food Policy 29(1): 15–44.

Neves, S. S., G. Swire-Clark, K. W. Hilu, and W. V. Baird. 2005. Phylogeny of Eleusine (Poaceae: Chloridoideae) based on nuclear ITS and plastid trnT-trnF sequences. Molecular Phylogenetics and Evolution 35: 395–419.

Nie, X., S. Lv, Y. Zhang, X. Du, L. Wang, S.S. Biradar, X. Tan, F. Wan, and S. Weining. 2012. Complete Chloroplast Genome Sequence of a Major Invasive Species, Crofton Weed (Ageratina adenophora). PLoS One 7(5): e36869.

Palmer, J. D. 1985. Evolution of chloroplast and mitochondrial DNA in plants and algae. pp. 131-240. In: Monograph in Evolutionary Biology: Molecular Evolutionary Genetics, Edited by R. J. MCINTYREP.l enum Press, New York.

Peng, Y., Z. Lai, T. Lane, M. Nageswara-Rao, M. Okada, M. Jasieniuk, H. O'Geen, R.W. Kim, R.D. Sammons, L.H. Rieseberg, and C.N. Stewart. 2014b. De Novo Genome Assembly of the Economically Important Weed Horseweed Using Integrated Data from Multiple Sequencing Platforms. Plant Physiol. 166(3): 1241–1254Available at http://www.plantphysiol.org/cgi/doi/10.1104/pp.114.247668.

Petit, C., G. Bay, F. Pernin, and C. Delye. 2010. Prevalence of cross-or multiple resistance to the acetyl-coenzyme A carboxylase inhibitors fenoxaprop, clodinafop and pinoxaden in

black-grass (Alopecurus myosuroides Huds.) in France. Pest management science, 66(2), 168-177.

Phillips, S. M. 1972. A survey of the Eleusine Gaertn. (Gramineae) in Africa. KewBull. 27: 251–270.

Pimentel, D., L. Lach, R. Zuniga, and D. Morrison. 2012. Environmental and Economic Costs of Nonindigenous Species in the United States. Calif. Biol. Sci. 50(1): 53–65.

Pop, M., and S. L. Salzberg. 2008. Bioinformatics challenges of new sequencing technology. Trends in Genetics, 24(3), 142-149.

Qian, J., J. Song, H. Gao, Y. Zhu, J. Xu, X. Pang, H. Yao, C. Sun, X. Li, C. Li, J. Liu, H. Xu, and S. Chen. 2013. The Complete Chloroplast Genome Sequence of the Medicinal Plant Salvia miltiorrhiza. PLoS One 8(2).

R Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing Vienna Austria.

Rahman, H., N. Jagadeeshselvam, R. Valarmathi, B. Sachin, R. Sasikala, N. Senthil, D. Sudhakar, S. Robin, and R. Muthurajan. 2014. Transcriptome analysis of salinity responsiveness in contrasting genotypes of finger millet (Eleusine coracana L.) through RNA-sequencing. Plant Mol. Biol. 85(4): 485–503.

Rambaut, A. 2009. FigTree. Tree figure drawing tool version 1.3. 1. Institute of Evolutionary biology, University of Edinburgh.

Randall, R. P. 2012. A global compendium of weeds. Department of Agriculture and Food Western Australia.

Raven, J. a, and J.F. Allen. 2003. Genomics and chloroplast evolution: what did cyanobacteria do for plants? Genome Biol. 4(3): 209.

Ravin, N. V., E. V Gruzdev, A. V Beletsky, A.M. Mazur, E.B. Prokhortchouk, M.A. Filyushin, E.Z. Kochieva, V. V Kadnikov, A. V Mardanov, and K.G. Skryabin. 2016. The loss of photosynthetic pathways in the plastid and nuclear genomes of the non- photosynthetic mycoheterotrophic eudicot Monotropa hypopitys. BMC Plant Biol. 16(Suppl 3)Available at http://dx.doi.org/10.1186/s12870-016-0929-7.

Reis-Filho, J. S. 2009. Next-generation sequencing. Breast Cancer Research, 11(3), S12.

Rice, P., I. Longden and A. Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16(6):26-277.

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R.

Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless and I. Birol. 2010. De novo assembly and analysis of RNA-seq data. Nature methods, 7(11), 909-912. doi:10.1038/10.1038/nmeth.1517

Ruby, J. G., P. Bellare, and J. L. DeRisi. 2013. PRICE: targeted assembly of components of (meta) genomic sequence data. G3 doi:10.1534/g3.113.005967.

Salse, Ã., B. Pie, R.C. Ã, and M. Delseny. 2004. New in silico insight into the synteny between rice (Oryza sativa L .) and maize (Zea mays L .) highlights reshuf ¯ ing and identi ® es new duplications in the rice genome. 2004: 396–409.

Schatz, M. C., J. Witkowski, and W. R. McCombie. 2012. Current challenges in de novo plant genome sequencing and assembly. Genome biology, 13(4), 243.

Shaw, J., E.B. Lickey, E.E. Schilling, and R.L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The Tortoise and the hare III. Am. J. Bot. 94(3): 275–288.

Shobana, S., K. Krishnaswamy, V. Sudha, N.G. Malleshi, R.M. Anjana, L. Palaniappan, and V. Mohan. 2013. Finger Millet (Ragi, Eleusine coracana L.): A Review of Its Nutritional Properties, Processing, and Plausible Health Benefits. 1st ed. Copyright © 2013 Elsevier Inc. All rights reserved.

Simão, F.A., R.M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E.M. Zdobnov. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19): 3210–3212.

Singh, P., and R. S. Raghuvanshi. 2012. Finger millet for food and nutritional security. African Journal of Food Science, 6(4): 77-84.

Singh, R.K., M. Lakshmi, V. Phanindra, V.K. Singh, A.U. Solanke, and P.A. Kumar. 2014. Isolation and characterization of drought responsive EcDehydrin7 gene from finger millet (Eleusine coracana (L.) Gaertn.). 74(4): 456–462.

Stamatakis, A. 2014. RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312-1313.

Stewart Jr, C. N. 2009. Weedy and invasive plant genomics. John Wiley & Sons.

Stewart Jr, C. N., P. J. Tranel, D. P. Horvath, J. V. Anderson, L. H. Rieseberg, J. H. Westwood, C. A. Mallory-Smith, M. L. Zapiola, and K. M. Dlugosch. 2009. Evolution of weediness and invasiveness: charting the course for weed genomics. Weed Science, 57(5), 451-462.

Tang, H., J. E. Bowers, X. Wang, R. Ming, M. Alam, and A. H. Paterson. 2008. Synteny and collinearity in plant genomes. Science, 320(5875), 486-488.

Tangphatsornruang, S., D. Sangsrakru, J. Chanprasert, P. Uthaipaisanwong, T. Yoocha, N. Jomchai, and S. Tragoonrung. 2010. The chloroplast genome sequence of mungbean (Vigna radiata) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Res. 17(1): 11–22.

Tehranchian, P., J.K. Norsworthy, M. Palhano, N.E. Korres, S. McElroy, H. Zhang, M. V. Bagavathiannan, and R.C. Scott. 2016. The Evidence for Reduced Glyphosate Efficacy on Acetolactate Synthase–Inhibiting Herbicide-Resistant Yellow Nutsedge (Cyperus esculentus). Weed Sci. 64(3): 389–398Available at https://www.cambridge.org/core/product/identifier/S0043174500021895/type/journal_article.

Thiel, T. 2003. MISA—Microsatellite identification tool.

VanBuren, R., D. Bryant, P.P. Edger, H. Tang, D. Burgess, D. Challabathula, K. Spittle, R. Hall, J. Gu, E. Lyons, M. Freeling, D. Bartels, B. Ten Hallers, A. Hastie, T.P. Michael, and T.C. Mockler. 2015. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature 527(7579): 508–511Available at http://www.nature.com/doifinder/10.1038/nature15714.

Vetrivel, U., V. Arunkumar, and S. Dorairaj. 2007. ACUA: a software tool for automated codon usage analysis. Bioinformation 2(2): 62–63.

Waterhouse, D. F. 1993. The major arthropod pests and weeds of agriculture in Southeast Asia. ACIAR.

Willis, J. C. 1973. A dictionary of the flowering plants and ferns. 8th edn. Cambridge University Press, Cambridge.

Wu, Z., and S. Ge. 2014. The whole chloroplast genome of wild rice (Oryza australiensis). Mitochondrial DNA (JUNE 2014): 1–2.

Wyman, S.K., R.K. Jansen, and J.L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20(17): 3252–3255.

Xu, Q., L.-L. Chen, X. Ruan, D. Chen, A. Zhu, C. Chen, D. Bertrand, W.-B. Jiao, B.-H. Hao, M.P. Lyon, J. Chen, S. Gao, F. Xing, H. Lan, J.-W. Chang, X. Ge, Y. Lei, Q. Hu, Y. Miao, L. Wang, S. Xiao, M.K. Biswas, W. Zeng, F. Guo, H. Cao, X. Yang, X.-W. Xu, Y.-J. Cheng, J. Xu, J.-H. Liu, O.J. Luo, Z. Tang, W.-W. Guo, H. Kuang, H.-Y. Zhang, M.L. Roose, N. Nagarajan, X.-X. Deng, and Y. Ruan. 2013. The draft genome of sweet orange (Citrus sinensis). Nat. Genet. 45(1): 59–66Available at http://www.ncbi.nlm.nih.gov/pubmed/23179022 (verified 28 January 2015).

Yi, D.K., and K.J. Kim. 2012. Complete chloroplast genome sequences of important oilseed crop Sesamum indicum L. PLoS One 7(5).

Yuan, J.S., P.J. Tranel, and C.N. Stewart. 2007. Non-target-site herbicide resistance: a family business. Trends Plant Sci. 12(1): 6–13.

Zerbino, D.R., and E. Birney. 2008. Velvet : Algorithms for de novo short read assembly using de Bruijn graphs. : 821–829.

Zhang, H., C. Li, H. Miao, and S. Xiong. 2013. Insights from the complete chloroplast genome into the evolution of Sesamum indicum L. PLoS One 8(11).

Zhang, H., N. Hall, J.S. McElroy, E.K. Lowe, and L.R. Goertzen. 2017. Complete plastid genome sequence of goosegrass (Eleusine indica) and comparison with other Poaceae. Gene 600: 36–43. Available at http://dx.doi.org/10.1016/j.gene.2016.11.038.

Zhao, Y., J. Yin, H. Guo, Y. Zhang, W. Xiao, C. Sun, J. Wu, X. Qu, J. Yu, X. Wang, and J. Xiao. 2015. The complete chloroplast genome provides insight into the evolution and polymorphism of Panax ginseng. Front. Plant Sci. 5(January): 1–12.

Zohary, D., and H. Meyer. 2000. Domestication of plants in the old world. Oxford, UK: Oxford University Press.