Phenotypic and genetic variation in an *Apios americana* breeding collection; and characterization of the HD-Zip gene family, involved in abiotic stress responses in *Glycine max*

by

Vikas Belamkar

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics and Genomics (Computational Molecular Biology)

Program of Study Committee:
Steven B. Cannon, Co-Major Professor
Thomas Lübberstedt, Co-Major Professor
Randy C. Shoemaker
Drena L. Dobbs
Mark E. Westgate

Iowa State University

Ames, Iowa

2015

DEDICATION

I dedicate this dissertation to my late grandfather P. S. Gajanana who inspired me to pursue higher education, and to my parents Prakash and Jayashree Belamkar, and Sourabha Shantappa without whose support and sacrifices, I would not have been able to complete this work.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

My sincere heartfelt thanks to my major advisor, Dr. Steven B. Cannon, for providing an opportunity to work in his lab, scientific guidance, independence in research, unwavering support, inspiring, and above all for being the finest human being with great values, and also one of the most calmest person during difficult times. I would also like to thank my co-major professor, Dr. Thomas Lübberstedt, for encouraging me to apply to Iowa State University and pursue my graduate studies, for scientific support, and career guidance. I'm also grateful to my other committee members: Dr. Randy C. Shoemaker, for providing laboratory facilities, assistance of people from his laboratory during RNA isolation and fieldwork, and professional guidance; and Dr. Drena L. Dobbs and Dr. Mark E. Westgate, for being helpful, motivating, and always available for discussions.

My special thanks to Dr. David M. Grant, Dr. Michelle A. Graham, and Dr N. Sathyanarayana, for their scientific guidance, critical discussions, and valuable advice.

A huge thank-you to Nathan T. Weeks for helping me with programming and coding, and to Scott R. Kalberer for assistance with phenotyping. Also, thanks to Ethalinda Cannon for helping in the field and for being supportive in the Apios research.

I would also like to acknowledge the efforts of collaborators: Dr. William J. Blackmon for sharing the Apios breeding collection, performing the field trial in Mechanicsville, VA, and inspiring all along; V. Gautam Bhattacharya for continuous encouragement, and sharing his unique and interesting perspectives on Apios research; Alex Wenger for performing the Apios field trial in Lititz, PA; and Andrew D. Farmer and Arvind K. Bharti for sequencing and bioinformatics analyses.

People who also deserve thanks are: Jody Hayes, Rebecca Nolan, Alex Gascho, and Joshua McCombs for their invaluable support during laboratory work, data collection and harvest; Ignacio Alvarez-Castro, Pedro Gonzalez and David Hessel for their guidance and for sharing ideas on statistical analyses of the phenotyping dataset.

Thanks also to members of the Cannon Lab and the SoyBase, Legume Information System and PeanutBase groups, and especially to Sudhansu Dash, Rex Nelson, Wei Huang, Kevin Feeley, Jacqueline Farrell and Jugpreet Singh for their friendship, discussions, and support that made me feel part of a wonderful family.

I also received help from Linda Wild in the Interdepartmental Genetics, Jaci Severson in Agronomy, and Leslie Elliott in the USDA Greenhouse.

My friends during graduate school, especially Ignacio Trucillo-Silva, Jackson Nteeba, Dinesh Thekkoot, Tao Zuo, Ratan Chopra, Bharath Narayana, Rajesha Rupaimoole, and Manoj Nair were supportive and encouraging all throughout.

A special thanks to Sourabha Shantappa for not only bearing my obsession on Apios, soybean, sequencing and my research work in general, but also being consistently supportive for the last 11 years, contributing to discussions, encouragement, time, meals, and for being extremely patient towards the completion of my graduate studies.

Lastly, I owe my family a great deal of gratitude for helping me throughout my life. My parents, Prakash and Jayashree Belamkar have made a lot of sacrifices, showed a great amount of patience, and helped me develop strength and courage to successfully pursue my career. Without their blessings, this work would never have been completed.

ABSTRACT

This dissertation has two main objectives: (1) morphological and genetic characterization of a little-studied edible legume native to North America, *Apios americana*; and (2) characterization of the soybean (*Glycine max*) homeodomain leucine zipper (HD-Zip) transcription factor family (involved in abiotic stress responses), and identification of candidate genes for dehydration and salt stress. In these projects, next generation sequencing (NGS) is evaluated as a tool for rapidly characterizing genetic variation (in Apios) and fine-scale genetic responses to abiotic stress (in soybean).

Apios, commonly called "potato bean," is a nitrogen-fixing legume that is adapted to diverse climatic conditions of central and eastern North America. It produces tubers (modified stem-tubers) that are rich in protein, have a long shelf life under refrigeration (>1 year), contain isoflavones, and have low levels of reducing sugars (potentially making the tubers useful for fried chips, for example). The plant was once a staple wild-collected food of Native American Indians, and is now a cultivated crop in Japan and South Korea. William J. Blackmon and Berthal D. Reynolds evaluated Apios as a new edible tuber crop in the US during the 1980s. Their breeding efforts during 1985-1994 lead to a collection of improved genotypes. As of 2010, 53 genotypes remained from Blackmon and Reynolds' work. As part of this dissertation project, phenotypic evaluation of these 53 genotypes was performed for multiple years, in multiple environments (Iowa, Virginia and Pennsylvania), and in three growing conditions (field, 305-mm [12-in.] pots and 381-mm [15 in.] grow-bags). Twenty traits were recorded, including 10 aboveground and 10 belowground measurements. There was significant variation among the genotypes for all but two emergence traits. Several

genotypes produced high yields - up to 1,515 g of tuber yield/plant. Transcriptome sequencing of multiple tissues from a single genotype generated both a high-quality *de novo* reference transcriptome assembly and an expression catalog. Re-sequencing of the leaf transcriptome from all the genotypes in the collection allowed identification of 58,154 high-quality SNPs and 39,609 gene expression markers (GEMs). Both SNPs and GEMs revealed population structure and pedigree relationships in the collection. Transcripts mapped to *Phaseolus vulgaris* (another legume in the Phaseoleae clade as Apios, with the same chromosome number and presumably similar genome structure) helped in building pseudo-Apios chromosomes. Linkage disequilibrium decay was investigated in the collection using putative genomic locations of the SNP markers, derived using the pseudo-Apios chromosomes. Association analysis conducted using SNPs and GEMs identified marker-trait associations for at least 11 traits. In summary, this study demonstrates accelerated and holistic (genomic and phenotypic) exploration of an underutilized crop.

The HD-Zip transcription factor family includes genes involved in abiotic stress. HD-Zip genes are well characterized in *Arabidopsis thaliana*, but not yet in soybean. As part of this dissertation project, HD-Zip genes were identified in the soybean genome using homology searches and Hidden Markov Model guided sequence alignments. Phylogeny reconstruction enabled placement of HD-Zip sequences into four previously described subfamilies. Syntenic paralog pairs were retained following polyploidy in *Glycine* ~13 Mya. RNA-Seq analysis identified 20 differentially expressed HD-Zip genes in the roots of soybean cv. 'Williams 82', at least at one of the three time points (1, 6, or 12 hr) under dehydration and salt stress. This indicates the role of HD-Zip genes in abiotic stress responses. Expression profiles generated for genes expressed in roots at 0, 1, 6 and 12 hr

under dehydration and salt stress will serve as an important resource for soybean genomic studies, and will aid in understanding plant responses to dehydration and salt stresses.

CHAPTER 1. INTRODUCTION

This dissertation has two main objectives: (1) morphological and genetic characterization of an *Apios americana* collection; (2) characterization of soybean (*Glycine max*) homeodomain leucine zipper (HD-Zip) transcription factor family, which is involved in abiotic stress responses, and identification of candidate genes for dehydration and salt stresses. Next generation sequencing (NGS) technologies provide tools to both explore new crops and more rapidly improve existing crops.

1.1 Need for Domestication of New and Climate-Resilient Crops

With an estimated 805 million hungry people in the world during 2012 - 2014, and climate change a reality, new and faster improvements are required to increase agricultural productivity. Great strides were made in increasing the world food production from ~2.94 billion metric tons (BT) in 1961 to ~8.27 BT in 2007[1]. This increase has been attributed to both improved cultivars and agronomic practices - with the amount of land available for farming remaining nearly constant over this time period[1]. However, increase in world food production needs to happen even amidst challenges such as climate change - resulting in extremes of low and high temperatures, decrease in available farmlands, and newly emerging biotic and abiotic stresses. Hence, improving climate-resilience along with improving genetics and management practices will be critical.

Fifteen crops provide 90% of the world's food energy intake, with three of them, rice, wheat, and maize, making up 60% (www.fao.org). Continuous improvement of these crops

will result in narrowing the genetic base of globally utilized crops, making food supply vulnerable to biotic and abiotic stresses. Increasing the diversity of staple crops may offer a hedge against failures of crops that span a limited range of agroecological niches. World Food Prize winner M. S. Swaminathan has been a strong advocate of biodiversity[2], which he refers to as "evergreen revolution." He has recently highlighted the importance of biodiversity and refers readers to important publications such as *Lost Crops of the Incas* and *Lost Crops of Africa*, which have documented the historic role of agrobiodiversity in ensuring food and health security[2]. Domestication of new crops is vital to increase agrobiodiversity.

1.2 Recent Success Stories of Domestication of New Crops

There are several examples of recent successes in domesticating new (underutilized, orphan, and non-staple) crops. Quinoa (*Chenopodium quinoa*) breeding has progressed rapidly in the last 4 to 5 years, along with wide acceptance in the society. Quinoa is described as a "complete" food because of the high protein content (~15%) and a well-balanced essential amino acid profile[3]. United Nations declared 2013 as International Year of Quinoa. Kevin Murphy and his team have developed a crossing method[4], evaluated texture differences among varieties of cooked Quinoa[5], identified cultivars tolerant to soil salinity[6], and performed extensive field evaluations to develop cultivars. Another success story is the exploration of perennial grains to reduce soil degradation and water contamination simultaneously[7]. DeHaan's research at The Land Institute to domesticate perennial intermediate wheat grass (Kernza™; *Thinopyrum intermedium*) has been fairly successful.

Two cycles of phenotypic selection, primarily based on seed yield per head and seed mass, evaluated across multiple locations and years, showed 77% increase in crop yield[8]. Palmer's grass (*Distichlis palmeri*), a halophytic (salt-tolerant) perennial crop produces wheat-like grains with a balanced amino acid profile, is gluten free and has a pleasant nutty flavor. The plant was a staple food of Cocopah Indians in the western United States, and is currently being explored as a promising crop in Australia (http://www.nypa.com.au/nypa-wild-wheat.html). Cactaceae, plants with Crassulacean Acid Metabolism (CAS) are highly drought tolerant, and species belonging to the genus Opuntia have been promising as a new vegetable crop[9, 10]. Legumes play a critical role in biological nitrogen fixation. In a recent study[11], an underutilized legume cover-crop *Mucuna pruriens* subsp. *utilis* used in rotation with corn, improved soil fertility and increased corn yield by nearly 60%. These examples clearly show that accelerated domestication of new crops is feasible.

1.3 Legumes for Addressing Food Security and Climate Change

The legume family comprises 670 to 750 genera and 18,000 to 19,000 species that include grain, forage, and agroforestry species[12, 13]. Grain legumes alone contribute nearly 33% of the dietary protein nitrogen (N) needs of humans[14]. The primary dietary legumes include (in rank order), bean (*Phaseolus vulgaris*), pea (*Pisum sativum*), chickpea (*Cicer arietinum*), broad bean (*Vicia faba*), pigeon pea (*Cajanus cajan*), cowpea (*Vigna unguiculata*), and lentil (*Lens esculenta*)[15]. In addition soybean and peanut (*Arachis hypogaea*) are also sources of dietary protein for the chicken and pork industries[13]. Both of

these legumes are responsible for more than 35% of the world's processed vegetable oil[13]. Clearly legumes play an important role in contributing to dietary needs of humans.

The legume family has traditionally been divided into three subfamilies: Papilionoideae, Mimosoideae and Caesalpinoideae[16] – though molecular phylogenies now divide the Caesalpinoideae into several taxa. Of the three subfamilies, the most extensively explored subfamily is Papilionoideae. This subfamily is further divided into two major clades (the millettoid/phaseoloid and the galegoid clade), and several smaller, early-diverging clades, including the genistoid and dalbergioid clades (containing lupin (*Lupinus* sp.) and peanut, respectively). The phaseoloid clade, popularly referred to as warm season legumes, contains many important crops: soybean, pigeon pea, common bean, mung bean (*Vigna radiata*), and cowpea, while the galegoid clade (cool season legumes) includes the model plants *Medicago truncatula* and *Lotus japonicus*, as well as alfalfa (*Medicago sativa*), chickpea, clover (*Trifolium* sp.), lentil, and garden pea. In the last five years, genomes of three warm season legumes (soybean, pigeon pea, and common bean)[17, 18, 19], and three cool season legumes (Medicago, Lotus, and chickpea)[20, 21, 22], and peanut (http://peanutbase.org) have been sequenced. Since the species within the Papilionoideae subfamily have high synteny across the species[23, 24, 25, 26], it is possible to translate insights gained from genome sequencing, along with available genomic resources to underexplored legume species. In this dissertation the genome sequence of different legumes especially common bean has been extensively utilized for genetic characterization of *Apios americana*. Genus Apios is an early-diverging lineage within the phaseoloid legumes, having separated from the remaining phaseoloids (including e.g. common bean and soybean) ~28 Mya[27, 28].

A hallmark of legumes is the ability of nearly 88% of the species to form nodules with rhizobia, and to fix atmospheric nitrogen in soil by symbiotic nitrogen fixation[29]. Generally legumes produce a podded fruit aboveground, and a few of them also produce tubers belowground. There are at least 54 species of legumes that produce tubers, and of them 27 species produce edible tubers[30]. To the best of our knowledge none of the tuber legumes have been extensively genetically characterized, and the physiology of tuber production in legumes is still unknown. Tuberous legumes are valuable to humans and animals because of their generally high protein and energy (carbohydrate) content. A few of the tuber producing legumes that are sparsely explored are winged bean (*Psophocarpus tetragonolobu*), Mexican yam bean (*Pachyrrhizus erosus*), kudzu (*Pueraria phaseoloides*), morama bean (*Tylosema esculentum*), and potato-bean (*Apios americana*).

In summary, the legume family comprises many species that are rich in micronutrients, are often resistance to drought, flooding and increased salinity, and can fix atmospheric nitrogen in the soil through symbiotic nitrogen fixation. Therefore, the legume family – with many species yet to be explored – appears to be encouraging for addressing food security and climate change.

1.4 Advances in Genomics as a Toolbox for Accelerating Improvement of Crops

Domestication of underutilized crops is particularly intriguing now because of the genomics revolution[31, 32]. NGS technologies can be utilized to generate genome or transcriptome assemblies (sequence information), expression catalogs, thousands of molecular markers - all at an affordable cost, even for a crop with absolutely no prior genetic

information[33]. Combining genetic information with the phenotypic data will further accelerate development of high yielding cultivars selected for useful traits[32]. Susan McCouch has advocated sequencing all non-duplicate samples in the world's gene banks, as a way to mine plant collections and better utilize biodiversity for food security[34]. Hence, NGS technologies have leveled the genomics field for both underutilized and staple crops, and procedures and methods developed to mine collections of underutilized crops can be helpful for staple crops as well.

In particular, the NGS technology that is extensively utilized in the present study is called RNA sequencing (RNA-Seq). RNA-Seq facilitates evaluation of the complete set of transcripts within a cell, in a high-throughput and quantitative manner. In this method, the complementary DNA (cDNA) produced from the mRNA of a specific tissue, or specific developmental stage or physiological condition, is sequenced on high-throughput sequencers. The sequencing data generated are then used to build a transcriptome assembly (reference sequence); further re-sequencing of the transcriptomes from other samples and aligning them to a reference allows identification of nucleotide variation (for, e.g., SNPs) in the transcripts; and sequences from different tissues mapped to a reference assembly will provide gene expression levels to develop an expression catalog[31, 32, 33, 35, 36]. Overall, RNA-Seq does not require knowledge of sequence information, and enables *de novo* reference development, expression profiling, and simultaneous marker discovery, marker validation and genotyping of the collection.

1.5 Goals of the Study

In this dissertation, new genomic tools have been evaluated to accelerate the characterization of a collection of *Apios americana*, and to thoroughly characterize the soybean HD-Zip gene family and identify candidate genes for dehydration and salt stress.

Apios americana has been an important plant to humans in North America for thousands of years, but following European colonization, has not been widely grown. This may be due to lack of awareness (and promotion) of Apios – and to lack of high-yielding cultivars, and to lack of markets for the crop. Dr. William J. Blackmon and Mr. Berthal D. Reynolds, in the 1980s, made the only significant effort toward improvement and popularization of Apios[37, 38]. Blackmon and Reynolds evaluated the food potential of Apios. They conducted a breeding program during 1985-1994 at Louisiana State University (LSU) Agricultural Experiment Station in Baton Rouge, LA. In their breeding program, germplasm was collected from the wild, and hybridized. Selections were then made to develop improved accessions. They identified several promising accessions. A cultivar LA85-034 produced high yields of tubers, and Blackmon and Reynolds considered this accession for release to smallholder farmers in the late 1980s. The breeding program ended in 1994 after Dr. Blackmon's retirement.

Blackmon and Reynolds made considerable progress in their breeding program. They demonstrated that rapid improvement of Apios is possible. This dissertation continues Blackmon and Reynolds' work. The primary goals of this dissertation are to (1) perform phenotypic evaluation of the 53 genotypes remaining from Blackmon and Reynolds's work - in multiple years, in multiple environments, and in three different growing conditions

(chapter 2); and (2) develop extensive genomic resources including, a high-quality reference *de novo* transcriptome assembly, an expression catalog utilizing six tissues, identify and genotype molecular markers using RNA-Seq, and perform genetic characterization of the collection (chapter 3). The ultimate goal of the Apios project is to combine the phenotypic and genomic data to support cultivar development.

The major goal of this study with respect to *Glycine max* (soybean) is to perform characterization of HD-Zip transcription factor family, which plays a significant role in abiotic stress responses. Candidate genes having roles in dehydration and salt stress are also identified (chapter 4).

In both the Apios and soybean projects, next generation sequencing is evaluated as a tool to rapidly explore new crops (e.g. Apios), and to improve cultivated crops (e.g. soybean).

1.6 References

1. Gustafson JP. Genomics and Breeding for Climate-Resilient Crops. In: *Vol. 1 Concepts and Strategies* (eds Kole C). Springer Berlin Heidelberg (2013).

2. Swaminathan MS. Gene Banks for a Warming Planet. *Science* **325**, 517 (2009).

3. Vega-Galvez A, Miranda M, Vergara J, Uribe E, Puente L, Martinez EA. Nutrition facts and functional potential of quinoa (Chenopodium quinoa willd.), an ancient Andean grain: a review. *J Sci Food Agric* **90**, 2541-2547 (2010).

4. Peterson A, Jacobsen S-E, Bonifacio A, Murphy K. A Crossing Method for Quinoa. *Sustainability* **7**, 3230-3243 (2015).

5. Wu G, Morris CF, Murphy KM. Evaluation of texture differences among varieties of cooked quinoa. *J Food Sci* **79**, S2337-2345 (2014).

6. Peterson A, Murphy K. Tolerance of Lowland Quinoa Cultivars to Sodium Chloride and Sodium Sulfate Salinity. *Crop Science* **55**, 331 (2015).

7. Cox TS, Van Tassel DL, Cox CM, DeHaan LR. Progress in breeding perennial grains. *Crop and Pasture Science* **61**, 513-521 (2010).

8. DeHaan LR. Progress in developing Kernza wheatgrass as a perennial grain. *Water, Food, Energy & Innovation for a Sustainable World*, (2013).

9. Segura S*, et al.* Genome sizes and ploidy levels in Mexican cactus pear species Opuntia (Tourn.) Mill. series Streptacanthae Britton et Rose, Leucotrichae DC., Heliabravoanae Scheinvar and Robustae Britton et Rose. *Genetic Resources and Crop Evolution* **54**, 1033-1041 (2007).

10. Mondragon Jacobo C. Cactus Pear Domestication and Breeding. In: *Plant Breeding Reviews* (eds Janick J). John Wiley & Sons, Inc. (2010).

11. Ortiz-Ceballos AI, Aguirre-Rivera JR, Salgado-Garcia S, Ortiz-Ceballos G. Maize–Velvet Bean Rotation in Summer and Winter Milpas: A Greener Technology. *Agron J* **107**, 330-336 (2015).

12. Raven P, Stirton C. Evolution and systematics of the Leguminosae. *Advances in legume systematics* **1**, 1-26 (1981).

13. Graham PH, Vance CP. Legumes: importance and constraints to greater use. *Plant Physiol* **131**, 872-877 (2003).

14. Vance CP, Graham PH, Allan DL. Biological Nitrogen Fixation: Phosphorus-A Critical Future Need? In: *Nitrogen fixation: From molecules to crop productivity* (eds Pedrosa FO, Hungria M, Yates G, Newton WE). Springer (2000).

15. *National Academy of Science (Biological Nitrogen Fixation)*. National Academy Press (1994).

16. Young ND, Bharti AK. Genome-Enabled Insights into Legume Biology. *Annual Review of Plant Biology* **63**, 283-305 (2012).

17. Schmutz J*, et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183 (2010).

18. Varshney RK*, et al.* Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat Biotech* **30**, 83-89 (2012).

19. Schmutz J*, et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics* **46**, 707-713 (2014).

20. Young ND*, et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-524 (2011).

21. Sato S*, et al.* Genome Structure of the Legume, Lotus japonicus. *DNA Research* **15**, 227-239 (2008).

22. Varshney RK*, et al.* Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat Biotech* **31**, 240-246 (2013).

23. Cannon SB*, et al.* Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular biology and evolution* **32**, 193-210 (2015).

24. Doyle J. Polyploidy in Legumes. In: *Polyploidy and Genome Evolution* (eds Soltis PS, Soltis DE). Springer Berlin Heidelberg (2012).

25. Lucas MR, Diop N-N, Wanamaker S, Ehlers JD, Roberts PA, Close TJ. Cowpea–Soybean Synteny Clarified through an Improved Genetic Map. *Plant Gen* **4**, 218-225 (2011).

26. Kang YJ*, et al.* Draft genome sequence of adzuki bean, Vigna angularis. *Sci Rep* **5**, (2015).

27. Li H*, et al.* Diversification of the phaseoloid legumes: effects of climate change, range expansion and habit shift. *Frontiers in plant science* **4**, 386 (2013).

28. Li J*, et al.* Phylogenetics and Biogeography ofApios(Fabaceae) Inferred from Sequences of Nuclear and Plastid Genes. *International Journal of Plant Sciences* **175**, 764-780 (2014).

29. De Faria SM, Lewis GP, Sprent JI, Sutherland JM. Occurrence of nodulation in the Leguminosae. *New Phytologist* **111**, 607-619 (1989).

30. Saxon EC. Tuberous Legumes: Preliminary Evaluation of Tropical Australian and Introduced Species as Fuel Crops. *Economic Botany* **35**, 163-173 (1981).

31. Edwards D, Batley J, Snowdon RJ. Accessing complex crop genomes with next-generation sequencing. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* **126**, 1-11 (2013).

32. Varshney RK, Terauchi R, McCouch SR. Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. *PLoS Biol* **12**, e1001883 (2014).

33. Ekblom R, Galindo J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity (Edinb)* **107**, 1-15 (2011).

34. McCouch S*, et al.* Agriculture: Feeding the future. *Nature* **499**, 23-24 (2013).

35. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics* **10**, 57-63 (2009).

36. De Wit P*, et al.* The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular ecology resources* **12**, 1058-1067 (2012).

37. Blackmon WJ, Reynolds BD. The crop potential of *Apios americana*-preliminary evaluations. *Hortscience* **21**, 1334-1336 (1986).

38. Reynolds BD, Blackmon WJ, Wickremesinhe E, Wells MH, Constantin RJ. Domestication of *Apios americana*. In: *Advances in new crops* (eds Janick J, Simon JE). Timber Press (1990).

CHAPTER 2. EVALUATION OF PHENOTYPIC VARIATION IN A COLLECTION OF *APIOS AMERICANA*: AN EDIBLE TUBEROUS LEGUME

A paper published in Crop Science 2015, 55(2):712-726

The electronic version of this article can be found online at:

https://www.crops.org/publications/cs/articles/55/2/712

Vikas Belamkar[ab], Alex Wenger[c], Scott R. Kalberer[d], V. Gautam Bhattacharya[e], William J. Blackmon[f], and Steven B. Cannon[bd]

a Interdepartmental Genetics, Iowa State Univ., Ames, IA 50011

b Dep. of Agronomy, Iowa State Univ., Ames, IA 50011

c 1529 Brunnerville Rd, Lititz, PA 17543

d USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011

e 532 Adam Ave, Ithaca, NY 14850

f 5097 Studley Rd, Mechanicsville, VA 23116

Authors' contributions

VB and SBC conceived and designed the experiments. AW performed the field trial and collected phenotype data in Lititz, PA; VB, SRK and SBC performed the field trial, conducted experiments in pots and grow-bags, and collected phenotype data in Ames, IA; WJB provided the accessions used in the study, and performed the field trial and collected

phenotype data in Mechanicsville, VA; VGB facilitated acquiring of germplasm used in the study; VB performed the data analyses; VB and SBC wrote the manuscript. All authors revised and approved the manuscript.

2.1 Abstract

Apios (*Apios americana* Medik.), sometimes called "potato bean," is a nitrogen-fixing legume, native to eastern North America, that produces protein-rich tubers at nodes along belowground stolons. Apios was used as a staple food source by Native Americans throughout eastern North America and holds promise as a crop. An Apios breeding program conducted during 1985 to 1994 involved the collection of ~210 wild accessions, followed by hybridization and selection, with assessments of >2200 lines. Of these, 53 genotypes were retained for further evaluation. The study reports the phenotypic variation in this collection, at three locations and under three growing conditions (field, pots, and grow-bags). We found significant ($P < 0.05$) variation among the genotypes for 18 of the 20 measured traits under field conditions, and for seven of 20 traits in pots and grow-bags. Internode length, plant vigor, and stem diameter at 2 and 5 mo had strong correlations ($r > 0.56$, $P < 0.01$) with belowground yield plant. Four phenotypically distinct clusters of genotypes were evident in the Apios collection. Several genotypes produced high yields in all locations—up to 1515 g of belowground tuber yield plant. The superior germplasm identified in this project may be suitable as cultivars, and will aid in further development of Apios lines as a crop.

2.2 Introduction

The human population relies on about 20 staple food crops, including cereals, tuber crops, legumes, sugar crops, coconuts (*Cocos nucifera* L.), and bananas (*Musa acuminata Colla*) (NAS, 1975). Reliance on a small number of plant species for food increases our vulnerability to catastrophic failures in the food system. Therefore, new crops are of great interest—particularly if they fill important niches either nutritionally or in agroecosystems. The edible tuber legume Apios (*Apios americana* Medik.) was once a staple wild-collected food source of Native American Indians (Beardsley, 1939; Blackmon and Reynolds, 1986). The characteristics that may make Apios valuable as a crop include: high nutritional value (with tubers rich in starch and protein) (Kikuta et al., 2011; Wilson et al., 1987); ease of cooking; good palatability; a long shelf life under refrigeration (>1 yr); adaptation across a wide geographical range in the United States and Canada (USDA NRCS, 2013); tolerance to a wide range of agricultural conditions-from well-drained loam to water-logged and acidic soils (Musgrave et al., 1991); and ability to fix atmospheric nitrogen through symbiosis with soil-resident rhizobial bacteria (Parker, 1999).

Apios is commonly called "potato bean," "Indian potato," "hopniss," and "American groundnut." It typically grows near creeks, rivers, and lakes. Apios produces both edible tubers and seeds, with tubers being of primary interest. The taste of tubers has been described as a mix of boiled peanut (*Arachis hypogaea* L.) and Irish potato (*Solanum tuberosum* L.) (Carlisi and Wollard, 2005). Tubers have low levels of reducing sugars (Carlisi and Wollard, 2005), and thus make excellent chips. Apios is now grown and used as a food crop in Japan and South Korea, because of its nutritional benefit relative to potatoes, sweet potatoes

[*Ipomoea batatas* (L.) Lam.], and taro [*Colocasia esculenta* (L.) Schott] (Kikuta et al., 2011; A. Wenger, unpublished data, 2013). The powder from the dried tubers is sold in markets as "apios powder," which is used as an ingredient in cookies, doughnuts, dumplings, and bread (Kikuta et al., 2011). Kikuta et al. (2011) compared the starch from Apios tubers, potatoes, maize (*Zea mays* L. ssp. *mays*), and Japanese arrowroot [*Pueraria lobata* (Lour.) Merr.], and found that Apios starch had properties similar to that of the arrowroot starch.

The Apios tubers are rich in proteins, carbohydrates, dietary fiber, and iron (Kikuta et al., 2011). Wilson et al. (1987) profiled the amino acids in the tubers and seed, and reported the crude protein to be between 25 and 30% for seeds, and 11 to 14% for tubers on dry defatted basis. Also, relative levels of essential amino acids were balanced both in the seeds and tubers, except for cysteine and methionine, which are usually low in legumes (Avraham et al., 2005). Apios tubers contain a novel isoflavone, genistein-7-O-gentiobioside, which is deglycosylated to synthesize genistein (Nara et al., 2011). These isoflavones have a potential role in radical scavenging and antioxidative activity (Nara et al., 2011; Takashima et al., 2013).

Screening for morphological variation requires a good understanding of the above- and belowground morphology of plants. The Apios vine varies from 1 to 6 m in length. The leaves are alternate, odd-pinnately compound, and have three to 11 leaflets (Fig. 1A). Apios produces stem/stolon tubers, like potato. Unlike potato, however, where only the terminal of the stolon tip enlarges to produce the tuber, Apios tubers are produced at most nodes along the stolon (Fig. 1B). A central "mother" (seed) tuber produces several bud meristems at one end. One or more meristems develop into the aboveground shoot or shoots, while the others develop into stolons- along which the "child tubers" are produced (Fig. 1B).

Apios has certain characteristics that makes it challenging for large-scale agriculture. It requires trellising, emerges late in the spring, and produces tubers that are often small (~0.02–0.04 m). For maximum yield, it needs to be harvested late in the fall, with considerable digging required to extract the tubers. Hence, breeding objectives should include increased yield and child-tuber size, decreased stolon length, and decreased vining tendency aboveground.

The initial requirement for development of improved cultivars is the collection and characterization of wild germplasm. The only substantial effort in this direction was by Blackmon and Reynolds (Blackmon and Reynolds, 1986; Reynolds et al., 1990), who collected nearly 210 wild Apios accessions from across the United States, with the majority of them sourced from Louisiana (150), Florida (14), and North Carolina (13). Blackmon and Reynolds used these in hybridization experiments from 1985 through 1994, and made preliminary evaluations of >2200 lines at the Louisiana State University Agricultural Experiment Station in Baton Rouge, LA. Several promising accessions were identified. An early cultivar, LA85-034, produced a high yield of tubers and was considered for release. Subsequent trials, on descendants of LA85-034 and other early lines, led to a collection of 53 accessions that remain from the Blackmon−Reynolds breeding work (Blackmon and Reynolds, 1986; Reynolds et al., 1990).

The germplasm collection, screening, and early breeding efforts during 1985 to 1994 constituted significant steps toward domestication of Apios. The collection used in this study comprised the 53 remaining genotypes from Blackmon and Reynolds' work, which had generally been selected for large child-tuber size, short stolon internodes, and uniform shape.

We report the first study involving evaluation of trait descriptors and characterization of phenotypic diversity in this Apios collection with the following objectives:

(i) Characterization of phenotypic variation in the collection, at three locations, and under three growing conditions (field, pots, and grow-bags).

(ii) Evaluation of 20 descriptors of above- and belowground traits to study the variability in the collection.

(iii) Identification of aboveground traits as markers for belowground tuber yield.

(iv) Identification of phenotypic clusters of genotypes in the collection.

(v) Identification of candidate parents and crossing schemes to develop cultivars with high yield and desired characteristics.

2.3 Materials and methods

2.3.1 Evaluation of Apios Collection under Field Conditions

The collection of 53 genotypes was grown at the North Central Regional Plant Introduction Station (NCRPIS), Ames, IA, in 2010 for seed tuber increase. During 2011 and 2012, the collection was evaluated in a randomized complete block design with two replications at NCRPIS, Ames, IA. In 2013 a subset of the collection was evaluated in two additional locations: 36 genotypes were evaluated in Mechanicsville, VA, and 20 genotypes in Lititz, PA. The genotypes were evaluated in multiple locations to investigate the effect of environmental variation on genotype performance, and to identify genotypes that have stable performances across different locations and in specific locations.

Individual genotypes were assigned to a single row of 0.9 m, composed of four tubers sown at a depth of ~0.06 m. The spacing between tubers within a row was ~0.3 m, and adjacent rows of genotypes were separated by 2.1 m. The soil characteristics and the weather conditions prevailing at the three locations during the experimental years are given in Supplemental Table S1. Apios plants were grown on trellises (1.5-m-high remesh wire panels). Water was mainly supplied by natural rainfall, with supplementary irrigation provided during extended dry periods. No inorganic fertilizers or herbicides were applied either before or during the experiment. Tubers were harvested after the growing season had ended and following a killing frost (late October in all 3 yr). Each genotype was harvested using a standard digging fork and a five-pronged broadfork, being careful to avoid breaking the stolons and preventing mixing of tubers belonging to different genotypes. The tubers were briefly air-dried and then stored in airtight plastic bags at 4°C.

2.3.2 Evaluation of Apios Collection in Pots and Grow-bags

The 53 genotypes of the collection were grown in 305-mm [12-in.] pots during 2011 and 2012, and 381 mm [15 in.] plastic grow-bags during 2012 and 2013 at a neighborhood farm, in Ames, IA (42°02′04.8″ N, 93°37′38.7″ W). Three plants per genotype were evaluated in both pots and grow-bags. The pots and bags were filled with soil to ~80% of capacity, and the plants were watered at least once a week, to maintain soil moisture. The characteristics of the soil used in pots and grow-bags are provided in Supplemental Table S1. The plants were harvested and processed at the end of growing season similarly to the field-grown plants.

2.3.3 Descriptors Used to Evaluate the Apios Collection

Based on the Bioversity International descriptors for tuber crops such as potato, yam (*Dioscorea* spp.), and sweet potato, 20 traits were selected to evaluate the collection (IBPGR/CEC, 1985, 1997; IBPGR/CIP/AVRDC, 1991; IPGRI/IITA, 1997) (Table 1). Ten traits were recorded on the aboveground part of the plant, and 10 belowground traits were measured after the plants were harvested.

During 2011 and 2012, the 20 traits were used to evaluate the collection grown at Ames, IA, under field conditions as well as in the pots and grow-bags. In 2013 the plants from the grow-bags were evaluated with 10 belowground measurements. Four selected belowground traits were used to evaluate plants grown in Mechanicsville, VA, and Lititz, PA (Supplemental Table S2).

2.3.4 Restricted Maximum Likelihood based Estimation of Variances

Restricted maximum likelihood (REML) (Patterson and Thompson, 1971) implemented in JMP Version 10 (SAS Institute Inc.) was used to estimate the variance components and the associated standard errors for traits. Restricted maximum likelihood analysis was performed on (i) 20 traits from the Ames, IA, trials conducted during 2011 and 2012; (ii) four yield-related traits recorded in Ames, IA, 2011 and 2012, and in 2013 in Mechanicsville, VA, and Lititz, PA—a total of four environments, 2 yr in IA, and 1 yr each in VA and PA; and (iii) 19 traits recorded on plants grown in pots and grow-bags, during 2011 to 2013 at Ames, IA.

For Analysis 1, year was treated as a fixed effect, whereas replicate within year, genotype, and genotype x year interaction were treated as random effects. Year was considered fixed because the experiment was conducted for only 2 yr, and years were not

randomly chosen (Piepho et al., 2003). For Analysis 2, environment was treated as a fixed effect, and replicate within environment, genotype, and genotype x environment interaction were treated as random effects. Environment was treated as a fixed effect because they were selected, that is, not randomly chosen, and the interest was to compute means for each environment, and compare the performances between environments (Piepho et al., 2003). For Analysis 3, REML was initially conducted separately on both pot and grow-bag data sets to determine variance component due to genotype under each of the conditions. Subsequently, REML analysis was conducted with treatment (pots vs. grow-bags) as fixed effect, and replicate within treatment, genotype, and genotype x treatment interaction as random effects. The phenotypic variance and broad-sense heritability for each of the traits were estimated as described in Fehr (1987). The REML-based least square (LS) means were calculated for each genotype and traits.

2.3.5 Multivariate Analyses

The multivariate analyses involving summary statistics, correlations and rank correlations, hierarchical clustering, and principal component analysis (PCA) were performed using REML- based LS means in R 2.15.2 (R Core Team, 2012). The description of R functions and packages used in the analyses are provided in the Supplemental Table S3. The top 10% of the highest performers in each of the environments were compared by means of Venn diagrams plotted using VENNY (Oliveros, 2007).

The genotypic diversity and the presence of subgroups in the collection were investigated using hierarchical cluster analysis and PCA. Both the cluster analysis and the PCA were performed on a scaled data set. The Euclidean distance matrix was generated using the transformed data set, and the intercluster distance was estimated with the Wards

(Ward, 1963) linkage method. A two-dimensional PCA plot of Principal Component (PC) 1 vs. PC2 was used to investigate phenotypic diversity.

2.4 Results and discussion

2.4.1 Variance Components and Broad-sense Heritability of Traits in Field Experiments

In the field trials conducted at Ames, IA, during 2011 and 2012, the genotype x year interaction was significant ($P < 0.05$) for one of the 20 traits (tuber-to-tuber distance) (Table 2). Hence, the mixed model without the genotype x year interaction was used to estimate variance components and LS means for each trait across both years (Bhargava et al., 2007). The year effect was significant ($P < 0.05$) for five aboveground traits (emergence time and first leaf emergence, leaflets recorded 2 and 5 mo after planting, and soil plant analysis development [SPAD]) and two belowground traits (yield plant and tuber-to-tuber distance).

The genotypic variance component was significant ($P < 0.05$) for 18 of the 20 traits, which indicated substantial amounts of phenotypic variation among the genotypes (Table 2). The variance component across the genotypes was nonsignificant for emergence time and the first leaf emergence. Differences were noticed among individual plants within a genotype for these two traits (data not shown). Similar results have been observed in the previous studies conducted by Blackmon and Reynolds (1986) and Reynolds et al. (1990). The reason for variation in emergence times among individual plants within a genotype is yet to be determined, and will be important to understand as production is scaled up.

The variance component for replicate within year was nonsignificant for all the traits (Table 2). The broad-sense heritability was highest (85%) for leaflets recorded 5 mo after

planting, and lowest (32%) for the first leaf emergence. Fourteen of the 20 traits had broad-sense heritability >61% (Table 2).

The environmental effect was significant ($P < 0.01$) for each of the four traits (yield plant, tubers plant, mother tuber weight, and child tuber weight) recorded at four environments (Ames, IA, 2011 and 2012, Mechanicsville, VA, and Lititz, PA) (Table 3). The genotypic variance was also significant ($P < 0.05$) for all four traits. The genotype x environment interaction was significant ($P < 0.01$) for three of the four traits (yield plant, tubers plant and mother tuber weight) and was not significant for child tuber weight. The broad-sense heritability ranged from 0.51 to 0.65 (Table 3). Similar results were obtained after excluding the relatively smaller data set from Lititz, PA, with the exception that the environmental effect for mother and child tuber weight was nonsignificant ($P < 0.05$) (data not shown).

To summarize, under field conditions we observed significant variation among the genotypes for all the traits evaluated except for emergence time and first leaf emergence. The high broad- sense heritability values for most of the traits demonstrated the repeatability of measurements across replicates, years, and environments, indicating the effectiveness of the phenotypic descriptors used in this study.

2.4.2 Variance Components and Broad-sense Heritability of Traits in Pots and Grow-bags Experiment

In the evaluations conducted in pots, the year effect was significant ($P < 0.05$) for 15 of the 19 traits (Supplemental Table S4). The four remaining traits (internode length, plant vigor, tubers plant, and child tuber length) did not have significant year effect. The genotypic variance was significant ($P < 0.05$) for three aboveground measurements (stem diameter

recorded 2 and 5 mo after planting, and leaflets recorded 2 mo after planting), and three belowground measurements (mother tuber weight, length, and width). The broad-sense heritability was >0.48 for six of the 19 traits.

In the experiment conducted in grow-bags, the year effect was significant ($P < 0.05$) for five traits (tubers plant, mother tuber weight, length, and width, and child tuber length). The rest of the five belowground measurements did not have significant year effect (Supplemental Table S4). The genotypic variance was significant ($P < 0.05$) for two belowground measurements: tubers plant and mother tuber length. The broad-sense heritability for tubers plant and mother tuber length was 0.64 and 0.45, respectively, and the rest of the traits had low broad-sense heritability values.

In the combined pots and grow-bags analysis, the treatment effect (pots vs. grow-bags) was significant ($P < 0.05$) only for yield plant (Table 4). The average yield in the grow-bags was 154 g, whereas in the pots it was 70 g. The genotypic variance was significant ($P < 0.05$) for four aboveground measurements (ground to first leaf, internode length, and leaflets recorded 2 and 5 mo after planting), and three belowground measurements (mother tuber weight, length, and width). The genotype x treatment interaction was not significant for each of the traits. The broad- sense heritability was >0.49 for seven of the 19 traits, whereas the rest of the traits had lower broad-sense heritability values.

In summary, the growth of genotypes in pots and grow-bags was not as uniform as under field conditions, and further experiments will be required to determine whether the growth of the Apios plants is limited in pots and grow-bags due to restricted space, nutrient limitations, or other stressors.

2.4.3 Summary Statistics of the Traits Evaluated on the Apios Collection

The summary statistics for 20 traits from the 2011 and 2012 Ames, IA, trials showed that tubers emerged 6 to 7 wk after planting, and the first leaf emerged a week later (Table 5). On average, there were five to seven leaflets per petiole. However, on individual plants within a given genotype, there were three to 11 leaflets (data not shown). The yield plant ranged from 183 to 537 g, with an average of 281 g. The number of tubers plant varied from 11 to 38, with an average of 19.

For the four traits evaluated at four environments, the mean as well as the maximum value was highest at Mechanicsville, VA, for yield plant (1057 g, 1515 g), tubers plant (56, 85), and mother tuber weight (184 g, 467 g), whereas the child tuber weight had highest average (49 g) and maximum (77 g) value in Ames, IA, in 2012 (Table 6).

In the field trials conducted in Ames, IA, during 2011 and 2012, Mechanicsville, VA, and Lititz, PA, the average yield plant was 174, 389, 1057, and 92 g, respectively. The top 10% of the high-yielding genotypes produced an average yield plant of 390, 627, 1393, and 146 g in these same four environments, respectively. Genotype 1972 was the highest yielding in all four environments, with yield plant values of 453, 681, 1515, and 164 g, respectively. The yields of the better-yielding lines are thus comparable to other tropical (sweet potato, cassava [*Manihot esculenta* Crantz], and yam [*Dioscorea*]) and temperate (potato) tuber and root crops (FAOSTAT, 2013). We believe the top 10% of the high-yielding genotypes in the four environments, and particularly the ones that performed well in more than one environment, are suitable for small-scale production. Apios may benefit resource-limited farmers, as it produces high yields of nutrient-dense food in small areas. Markets and

increased public acceptance for this promising species will follow exposure to the crop, particularly as market gardeners gain access to higher-yielding varieties.

2.4.4 Phenotypic Correlations among the Above- and Belowground Traits

In the 2011 and 2012 Ames, IA, field trials, leaflets measured 2 mo after planting had high positive associations ($r > 0.57$, $P < 0.01$) with the three mother tuber measurements (weight, length, and width), suggesting that the mother tuber plays a vital role in early stages of plant growth (Fig. 2 and Supplemental Table S5).

The four aboveground traits (internode length, plant vigor, and stem diameter measured at 2 and 5 mo after planting) had high positive correlations ($r > 0.56$, $P < 0.01$) with yield plant. In addition, internode length, plant vigor, and stem diameter measured 5 mo after planting were also highly correlated ($r > 0.51$, $P < 0.01$) with the three child tuber measurements (weight, length, and width), and the mother tuber length. The stem diameter measured 2 mo after planting had a positive association ($r > 0.50$, $P < 0.01$) with two mother tuber measurements (weight and length), three child tuber measurements (weight, length, and width), tuber-to-tuber distance, and stolon length. The stem diameter measured 5 mo after planting also had a positive association ($r > 0.53$, $P < 0.01$) with tuber-to-tuber distance.

Similar correlations are known in other tuber and root crops, such as a positive correlation between internode length and tuber yield per hectare in sweet potato (Choudhary et al., 2000). Larger stem diameter corresponding to higher yield is observed in cassava (Sankaran et al., 2008), and higher plant vigor is associated with higher tuber yield in potato (Golmirzaie and Ortiz, 2002; Ortiz and M. Golmirzaie, 2003).

Identification of aboveground traits that can serve as a proxy for belowground yield will aid in early identification of high-yielding genotypes. In the current evaluations, stem

diameters were positively associated with both desirable (yield plant) and undesirable (stolon length and tuber-to-tuber distance) traits. Therefore, internode length and plant vigor, both of which have strong phenotypic correlation and also genetic correlation (0.71 and 0.77 [data not shown]) with the belowground yield, can be used as proxy traits for belowground yield.

Plant vigor, recorded 5 mo after planting, may provide an indication of high-yielding plants. However, it would be useful to test whether the plant vigor recorded at the beginning of the growing season could also be highly correlated with yield. The relationship between early plant vigor and high yield has been well established in potato (Golmirzaie and Ortiz, 2002; Ortiz and M. Golmirzaie, 2003). This will eventually aid in short-listing two traits that can be used during emergence and elongation of the vine, for selection of high-yielding plants.

2.4.5 Phenotypic Correlations among the Four Traits Recorded in Four Environments

The correlation analysis between the same traits across the four environments yields six pairwise comparisons for each trait. The yield plant and tubers plant were significant ($P < 0.05$) in five of the six pairwise comparisons, whereas the mother tuber and child tuber weights had all six pairwise correlations statistically significant (Table 7). The child tuber weight correlation coefficients ranged from 0.95 to 0.97, which indicates relatively similar performances of the genotypes in the collection for child tuber weight across all four environments.

Twelve of the 16 pairwise comparisons generated from correlations between mother tuber weight and tubers plant in individual environment and across the four environments are significantly ($P < 0.05$) negatively correlated, which suggests that plants with large mother tubers make fewer child tubers. Similarly, the yield plant was significantly ($P < 0.05$)

correlated with child tuber weight in 14 of the 16 pairwise comparisons. Thus, yield plant and child tuber weight are strongly correlated within as well as across environments.

Evaluation of Apios germplasm in four environments indicated the use of the trait "child tuber weight" as a secondary trait for selecting high-yielding Apios genotypes. The child tuber weight and yield plant are strongly positively correlated within and across environments, and the genotype x environment interaction is not significant for child tuber weight. The broad-sense heritability of child tuber weight (0.65) is higher than that of the yield plant (0.51). Hence, based on the results obtained in this study, the selection of Apios genotypes for higher yield can be efficiently pursued by selecting genotypes with higher child tuber weight.

2.4.6 Rank Correlations of the Four Traits Recorded in Four Environments

The estimation of rank correlations and the comparison of top 10% of the performers for yield plant, tubers plant, mother tuber weight, and the child tuber weight recorded in four environments identifies genotypes with stable performances across the environments. The rank correlation was significant ($P < 0.05$, $r > 0.52$) for all four traits measured in Ames, IA, 2011 and 2012 (Table 8). The mother tuber weight had significant ($P < 0.05$) rank correlation between measurements made in Ames, IA, 2011 and Mechanicsville, VA ($r = 0.37$), and Ames, IA, 2012 and Mechanicsville, VA ($r = 0.30$). Similarly, the child tuber weight had significant rank correlation ($r = 0.29$) between measurements made in Ames, IA, 2012 and Mechanicsville, VA.

The genotypes in the top 10% based on their performances for traits at each of the four environments are shown in Fig. 3. Genotype 1972 was the highest yielding in all four environments, and it also had highest values for child tuber weight in all four environments.

Genotypes 807 and 898 were among top 10% for tubers plant in three environments each. Genotype 1661 was in the top 10% for mother tuber weight in three environments. Genotype 1849 was in the top 10% for child tuber weight in all four environments. Thus, genotypes that occur in the top 10% have stable performances across different environments, with a few exceptions. For instance, Genotype 2155 and 2201 were among the top 10% for yield plant only at Ames, IA, in 2012. In short, the Apios collection contained genotypes that performed in a stable manner across all environments and some that excelled only in specific environments.

2.4.7 Hierarchical Clustering Analysis

The hierarchical clustering analysis on data from the 2011 and 2012 Ames, IA, trials grouped the genotypes based on similarities in the performance of the traits recorded. The collection was subdivided into two phenotypic subgroups (Fig. 4). Subgroup A mainly included genotypes that emerged early and had high values for most of the traits recorded, whereas subgroup B contained genotypes that emerged late and had low values for most of the traits measured. Subgroups A and B were further subdivided into two clusters each and had biological basis (Fig. 5 and 6). Clusters A1 and A2 separated the high-performing genotypes into groups based on belowground growth patterns. Cluster A1 contained genotypes that produced large mother tubers with short stolons and few child tubers (Fig. 5 and 6), while Cluster A2 comprised genotypes that had high yield plant (stolons and tubers), and also had the highest mean values for all belowground traits except for the mother tuber–related measurements (Fig. 5 and 6). The mean of the Cluster A1 had highest value for three aboveground traits, leaflets measured 2 and 5 mo after planting, and SPAD, while Cluster A2 had highest average values for five aboveground traits (ground to first leaf, internode length,

stem diameter measured 2 and 5 mo after planting, and plant vigor), and lowest value for SPAD. The Clusters B1 and B2 mainly represented intermediate and poorly performing genotypes, respectively. Cluster B2 had the smallest mean for all the measurements except for three traits (SPAD, tubers plant, and stolon length), whereas cluster B1 had intermediate mean value for all but six traits (ground to first leaf, SPAD, tubers plant, tuber-to- tuber distance, stolon length, and child tuber length). Overall, the cluster analysis identified four distinct clusters in the Apios collection.

2.4.8 Principal Component Analysis

The PCA was used to reduce the dimensionality by identifying traits that adequately explained most of the phenotypic variation observed at Ames, IA, during 2011 and 2012. We analyzed four PCs that explained a total of 74% of the variation (Table 9). The PC1 was comprised primarily of the internode length, stem diameter measured 2 and 5 mo after planting, plant vigor, yield plant, and child tuber weight, length, and width. These eight traits represent high-yielding Apios genotypes. The PC2 had negative contributions from four traits: leaflets measured 2 mo after planting, the three mother tuber–related measurements (weight, length, and width); and had positive contributions from two traits: tubers plant and the stolon length. The PC2 is, interestingly, a contrast between the two types of high-performing groups based on the distinct belowground growth patterns. The traits that contributed to PC2 represented the characteristics of the genotypes that make bigger mother tubers, shorter stolons, and a smaller number of child tubers. The PC3 had major contributions from emergence time and first leaf emergence, whereas the PC4 had contributions from leaflets measured 5 mo after planting and tuber-to-tuber distance. In

summary, based on the PCA results, 14 traits explained most of the variation observed (59%), and can be effectively utilized in multienvironment trials.

Principal component analysis was also used to investigate the phenotypic diversity in the collection. A plot of PC1 vs. PC2 divided the collection into four groups representing two generally high-performing groups (i.e., highest average values for most of the traits), an intermediate-, and a poorly performing group (Fig. 7). These results agree with the cluster analysis, with the exception of two genotypes. Genotype 1587 has a higher yield plant and thus belongs to a cluster of high-yielding genotypes as classified by cluster analysis; yet most of the other traits, such as internode length, stem diameter recorded 2 and 5 mo after planting, and child tuber weight, classify this genotype in the intermediate group in accordance with PCA. Similarly, Genotype 1846 has a mixed response for most of the traits, yielding conflicting PCA and clustering results.

To summarize, both PCA and the clustering analysis identified four clusters in the collection with distinct phenotypes, thereby providing clues for the selection of suitable genotypes with desired phenotypes, and suggesting future crossing schemes.

2.4.9 Candidate Genotypes and Potential Crossing Schemes for Developing the First Cultivars of Apios

The prime objective of our Apios breeding program is to develop high-yielding cultivars with desirable traits: large child tubers, shorter stolons, and lower tuber-to-tuber spacing. The top 10% of high-yielding genotypes at one or multiple environments investigated in this study are genotypes 1972, 2191, 898, 2127, 1849, 2155, 2201, 1970, and 2065 (Fig. 3A). These may serve as a source of parents for developing higher-yielding cultivars. A common standard procedure for improving yield plant beyond the current level,

is by making a “good x good” cross (Fehr, 1987). However, it is important to select genetically diverse genotypes to retain the genetic variability in the high-yielding cultivar. Apios may exhibit hybrid vigor. This would be apparent as a positive relationship between the yield of the plant and its level of heterozygosity. This association has been previously observed in other clonally propagated heterozygous plant species, including potato (Fehr, 1987; Ortiz and M. Golmirzaie, 2003). High levels of heterozygosity were found in a set of eight randomly chosen genotypes from the Apios collection using single nucleotide polymorphism markers (Cannon and Belamkar, unpublished data, 2013). Clonal propagation can be taken advantage of to produce increases of desirable heterozygous lines. A heterozygous cultivar x cultivar cross results in a segregating population, and the seeds from this cross can be planted, giving new clonal lines that can be screened for high-value lines, which can then be maintained as new cultivars (Fehr, 1987).

The shorter-stolon trait, with intermediate tuber-to-tuber distance, is characteristic of the genotypes that belong to the first high-performing cluster (Fig. 4 and 5). A plausible strategy for breeding a heterozygous population of high-yielding genotypes with shorter stolon length and lower tuber-to-tuber spacing would be to make crosses of genotypes from the first cluster as donor parent (1661, 2170, 1908, 2179, 2148, 2010) with high-yielding genotypes (1972, 2191, 898, 2127, 1849, 2155, 2201, 1970, 2065) (Fig. 3A) as recurrent parents, using a modified backcross for clonally propagated heterozygous plant species (Fehr, 1987). Similarly, the use of genotype 807 or 898, which consistently produce a large number of smaller child tubers (Fig. 3B) as the donor parent and high-yielding genotypes as recurrent parents, may result in high- yielding genotypes with a large number of bigger child tubers.

Apios has long been an important wild food source and has potential as a new crop (Beardsley, 1939; Reynolds et al., 1990). In an era of global food insecurity and climate change, potential new crop plants have heightened value. Exceptional qualities - including the ability to fix nitrogen, reasonable resistance toward biotic and abiotic stresses, high protein content, long shelf life of the tubers, potential for medicinal applications, a highly diverse collection, and existence of high-yielding genotypes - all make Apios a strong candidate for continued work toward domestication, and use as a novel crop.

Supplemental Information Available

Supplemental Tables S1-S5 provides the following information:

(1) Soil characteristics and the weather conditions at Ames, IA during 2011-2012, and Mechanicsville, VA and Lititz, PA in 2013.

(2) Phenotypic data associated with each of the experiments described in this study.

(3) R functions and packages used for statistical analyses.

(4) Variance estimates and broad-sense heritability estimated separately for pot and grow-bag experiments.

(5) Correlation coefficients with significance values among the 20 traits measured on the Apios collection in Ames, IA 2011-2012.

2.5 Acknowledgments

The authors are thankful to the late Berthal D. Reynolds for his contribution to the pioneering germplasm collection and breeding efforts in the 1980s and 1990s; to Dr. Dennis Wollard for providing seed tubers of a subset of the genotypes in the Apios collection; to Dr. Randy C. Shoemaker for providing laboratory facilities as well as consistent encouragement; and to Jody Hayes, Rebecca Nolan, Alex Gascho, and Joshua McCombs for their invaluable support during data collection and harvest.

2.6 References

1. Avraham, T., H. Badani, S. Galili, and R. Amir. 2005. Enhanced levels of methionine and cysteine in transgenic alfalfa (*Medicago sativa L.*) plants over-expressing the Arabidopsis cystathionine γ-synthase gene. Plant Biotechnol. J. 3:71-79. doi:10.1111/j.1467-7652.2004.00102.x.

2. Beardsley, G. 1939. The groundnut as used by the Indians of eastern North America. Papers of the Michigan Academy of Science, Arts and Letters 25:507-525.

3. Bhargava, A., S. Shukla, and D. Ohri. 2007. Genetic variability and interrelationship among various morphological and quality traits in quinoa (*Chenopodium quinoa* Willd.). Field Crops Res. 101:104-116. doi:10.1016/j.fcr.2006.10.001.

4. Blackmon, W.J., and B.D. Reynolds. 1986. The crop potential of *Apios americana* - Preliminary Evaluations. HortScience 21:1334-1336.

5. Carlisi, J., and D. Wollard. 2005. History, Culture, and Nutrition of *Apios americana*. J. Nutraceut. Funct. Med. Foods 4:85-92.

6. Choudhary, S.C., H. Kumar, V.S. Verma, and S.K.T. Nasar. 2000. Correlation and path analysis in potato (*Ipomoea batatas* L.). Journal of Research, Birsa Agricultural University 12:239-242.

7. FAOSTAT. 2013. Food and agricultural organization of the United Nations. http://faostat.fao.org/site/567/default.aspx#ancor (accessed 16 Apr. 2013).

8. Fehr, W.R. 1987. Principals of cultivar development, vol. 1. Theory and TechniqueMcGraw-Hill Inc, New York.

9. Golmirzaie, A., and R. Ortiz. 2002. Inbreeding and true seed in tetrasomic potato. III. Early selection for seedling vigor in open-pollinated populations. Theor. Appl. Genet. 104:157-160. doi:10.1007/s001220200019.

10. IBPGR/CEC. 1985. Minimum list of characteristics of potato varieties. G.R. Mackay, M.J. Hijink, G. Mix, editors. Commission of European communities (CEC) secretariat, Brussels/International Plant Genetic Resources Institute, Rome, Italy.

11. IBPGR/CEC. 1997. Descriptors for the cultivated potato. Z. Huamán, J.T. Williams, W. Salhuana and L. Vincent, editors. International Board for Plant Genetic Resources, Rome Italy.

12. IBPGR/CIP/AVRDC. 1991. Descriptors for Sweet Potato. Huamán, Z., editor. International Board for Plant Genetic Resources, Rome Italy.

13. IPGRI/IITA. 1997. Descriptors for Yam (*Dioscorea* spp.). International Institute of Tropical Agriculture, Ibadan, Nigeria/International Plant Genetic Resources Institute, Rome, Italy.

14. Kikuta, C., Y. Sugimoto, Y. Konishi, Y. Ono, M. Tanaka, K. Iwaki, et al. 2011. Physicochemical and Structural Properties of Starch Isolated from *Apios americana* Medikus. Journal of Applied Glycoscience 59:21-30. doi:10.5458/jag.jag.JAG-2011_011.

15. Musgrave, M.E., A.G. Hopkins Jr, and C.J. Daugherty. 1991. Oxygen insensitivity of photosynthesis by waterlogged *Apios americana*. Environ. Exp. Bot. 31:117-124. doi:http://dx.doi.org/10.1016/0098-8472(91)90014-F.

16. Nara, K., K.-i. Nihei, Y. Ogasawara, H. Koga, and Y. Kato. 2011. Novel isoflavone diglycoside in groundnut (*Apios americana* Medik). Food Chem. 124:703-710. doi:10.1016/j.foodchem.2010.05.107.

17. NAS. 1975. Underexploited tropical plants with promising economic value, Washington, D.C.

18. Oliveros, J.C. 2007. VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html (accessed 7 Feb. 2014).

19. Ortiz, R., and A.L.I. M. Golmirzaie. 2003. Genetic parameters for agronomic characteristics. I. Early and intermediate breeding populations of true potato seed. Hereditas 139:212-216. doi:10.1111/j.1601-5223.2003.01734.x.

20. Parker, M.A. 1999. Relationships of Bradyrhizobia from the Legumes *Apios americana* and *Desmodium glutinosum*. Appl. Environ. Microbiol. 65:4914-4920.

21. Patterson, H.D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545-554. doi:10.1093/biomet/58.3.545.

22. Piepho, H.P., A. Büchse, and K. Emrich. 2003. A Hitchhiker's Guide to Mixed Models for Randomized Experiments. Journal of Agronomy and Crop Science 189:310-322. doi:10.1046/j.1439-037X.2003.00049.x.

23. R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/ (accessed 12 Nov. 2013).

24. Reynolds, B.D., W.J. Blackmon, E. Wickremesinhe, M.H. Wells, and R.J. Constantin. 1990. Domestication of *Apios americana*. In: J. Janick and J. E. Simon, editors, Advances in new crops. Timber Press, Portland, OR. p. 436-442.

25. Sankaran, M., N.P. Singh, C. Datt, B. Santhosh, M. Nedunchezhiyan, S.K. Naskar, et al. 2008. Evaluation of High Yielding Cassava Varieties under Upland Conditions of Tripura. Journal of Root Crops 34:73-76.

26. Takashima, M., K. Nara, E. Niki, Y. Yoshida, Y. Hagihara, M. Stowe, et al. 2013. Evaluation of biological activities of a groundnut (*Apios americana* Medik) extract containing a novel isoflavone. Food Chem. 138:298-305. doi:10.1016/j.foodchem.2012.10.100.

27. USDA NRCS. 2013. The PLANTS Database. http://plants.usda.gov/java/profile?symbol=APAM (accessed 7 Mar. 2013).

28. Ward, J.H. 1963. Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58:236-244. doi:10.1080/01621459.1963.10500845.

29. Wilson, P.W., F.J. Pichardo, J.A. Liuzzo, W.J. Blackmon, and B.D. Reynolds. 1987. Amino Acids in the American Groundnut (*Apios americana*). J. Food Sci. 52:224-225. doi:10.1111/j.1365-2621.1987.tb14013.x.

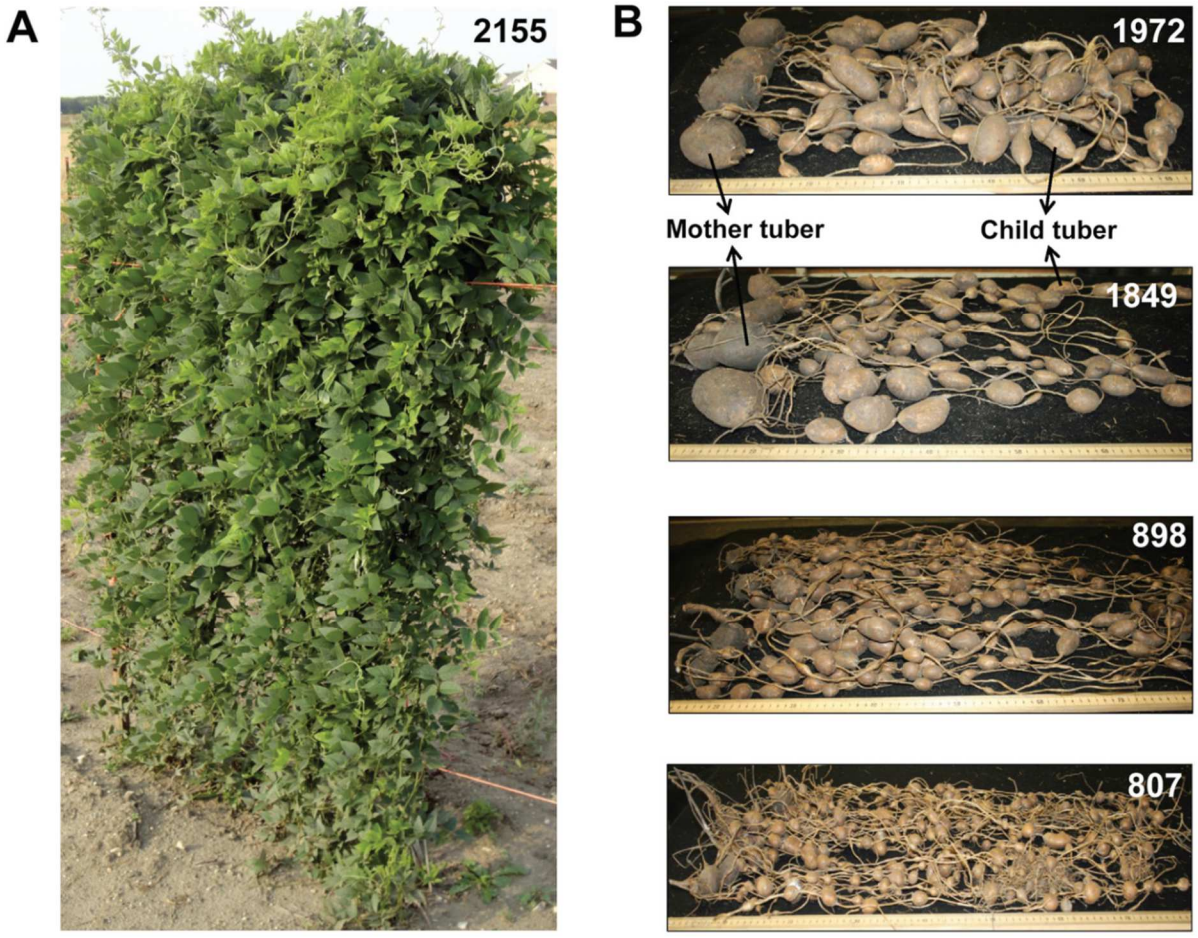

Figure 1. The morphology of field grown Apios plants in 2012 (A) Above-ground morphology of genotype 2155 - five months after planting at North Central Regional Plant Introduction station (NCRPIS) in Ames, IA (B) Below-ground morphology showing tubers and stolons of four Apios genotypes (1972, 1849, 898 and 807) after harvest at Ames, IA.

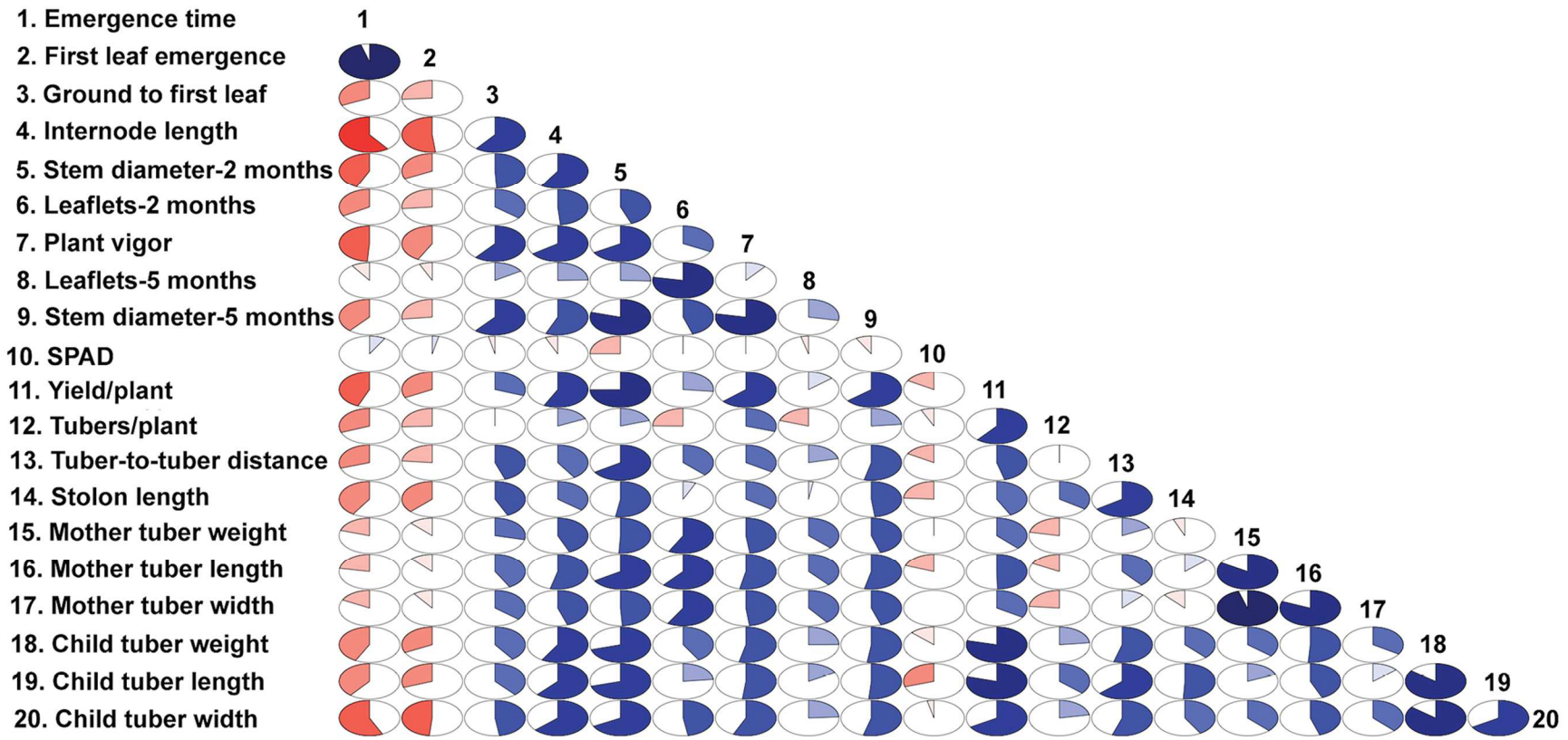

Figure 2. Correlation coefficients between pairs of 20 above and below-ground traits measured on the Apios collection grown at Ames, IA during 2011 and 2012. Negative correlations are red; positive values are blue. Values close to zero are lightly colored.

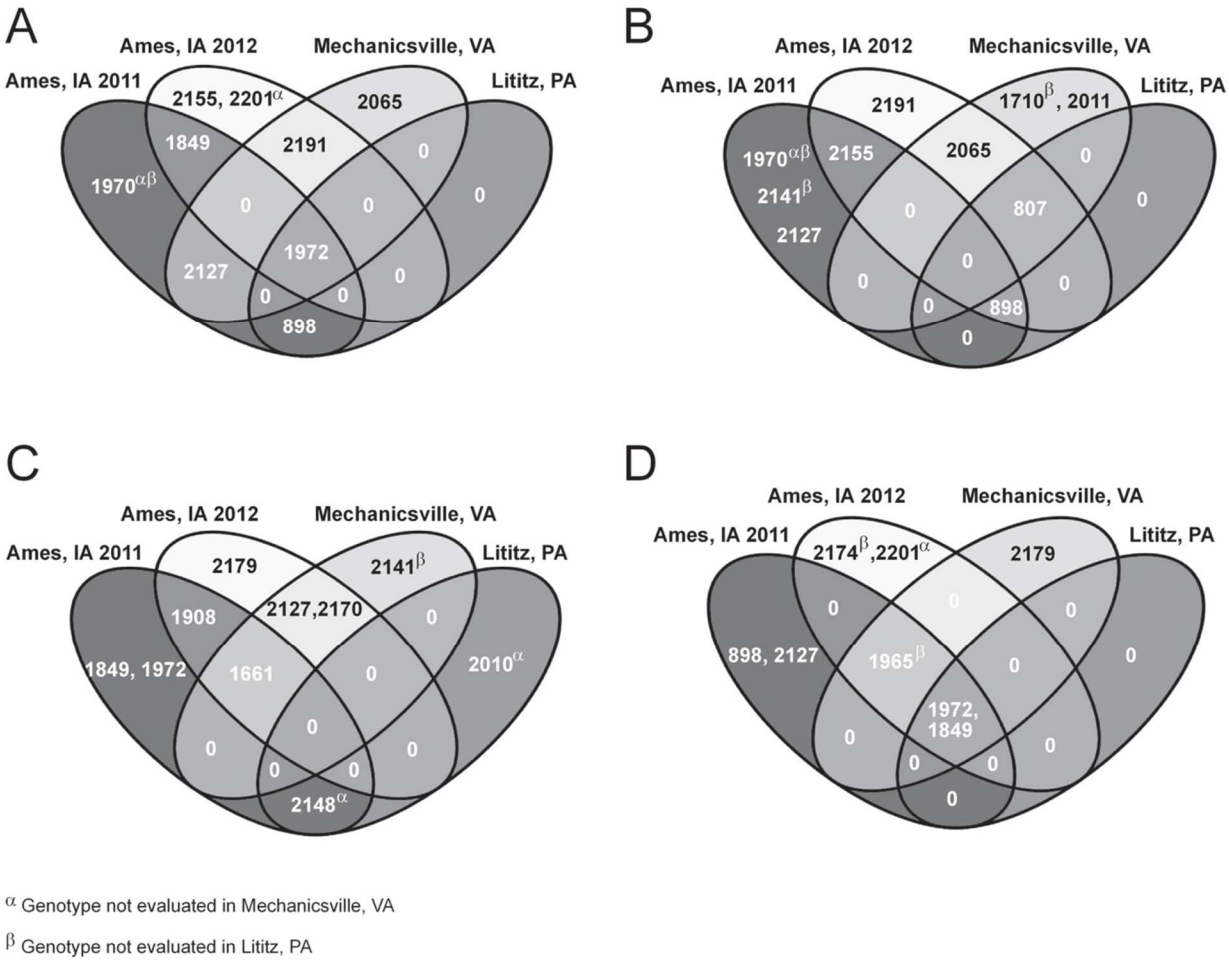

Figure 3. Comparison of top 10% of the Apios genotypes for the four traits evaluated in Ames, IA during 2011 and 2012, Mechanicsville, VA and Lititz, PA in 2013 (A) Yield/plant (B) Tubers/plant (C) Mother tuber weight (D) Child tuber weight.

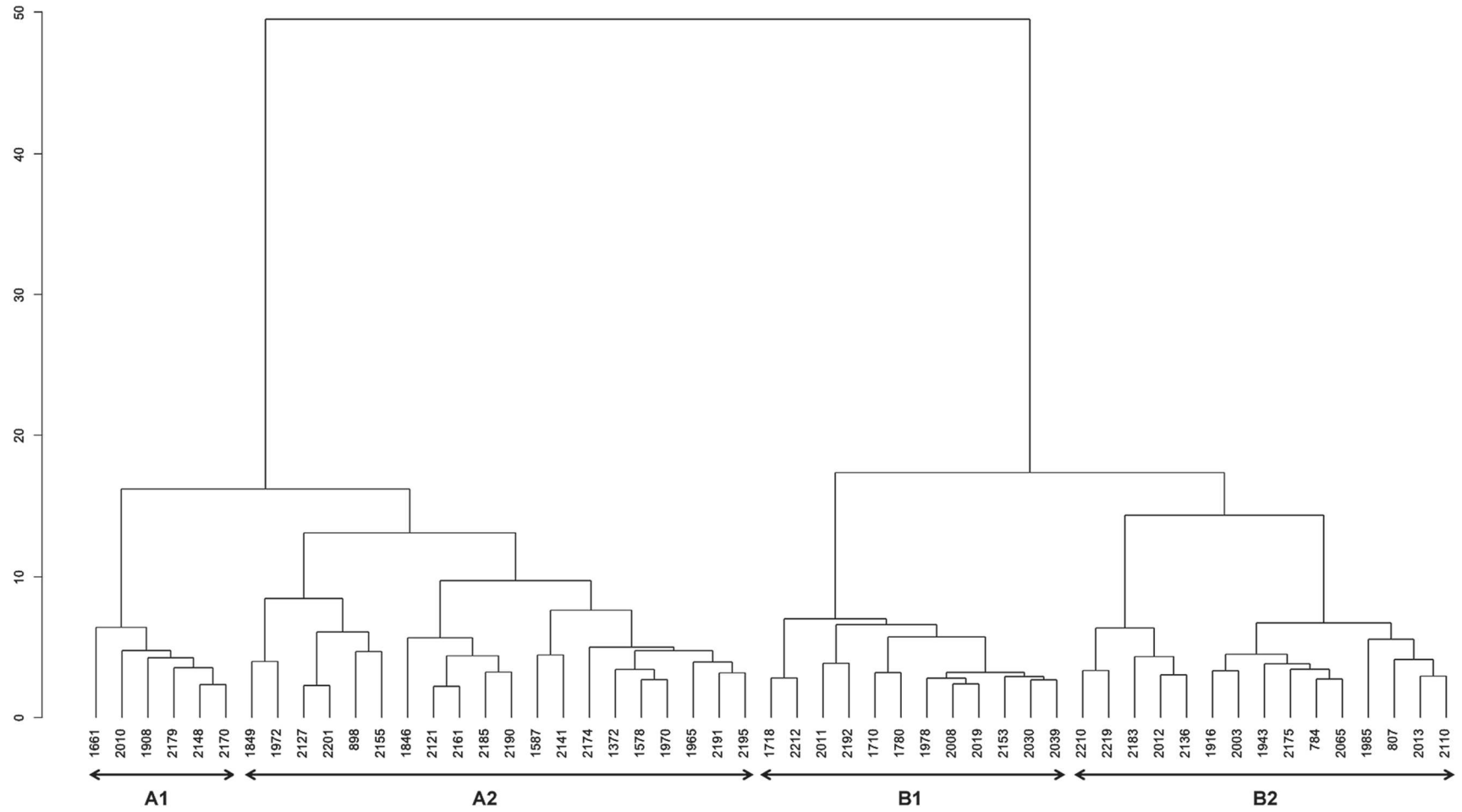

Figure 4. A dendrogram showing clusters of genotypes in the Apios collection comprising 53 genotypes evaluated at Ames, IA during 2011 and 2012. The clusters A1 and A2 represent high-performing genotypes, while B1 and B2 represent intermediate and poorly-performing genotypes, respectively.

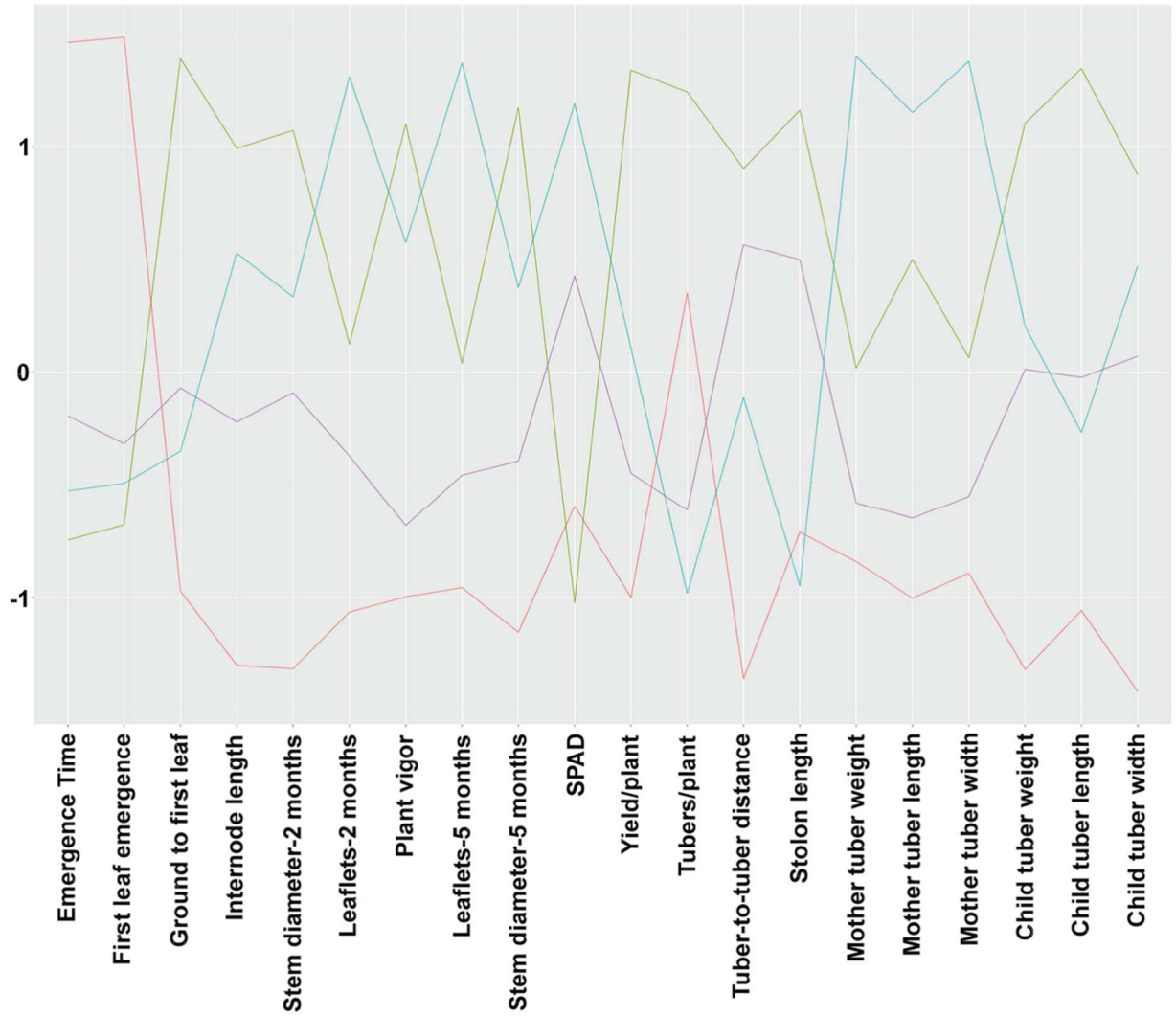

Figure 5. A parallel coordinate plot of the four clusters-means for each trait measured on the Apios collection at Ames, IA during 2011 and 2012. Clusters A1, A2, B1 and B2 from Figure 4 are shown in blue, green, purple and red respectively. Each colored line shows average response of genotypes of a cluster, for each of the traits.

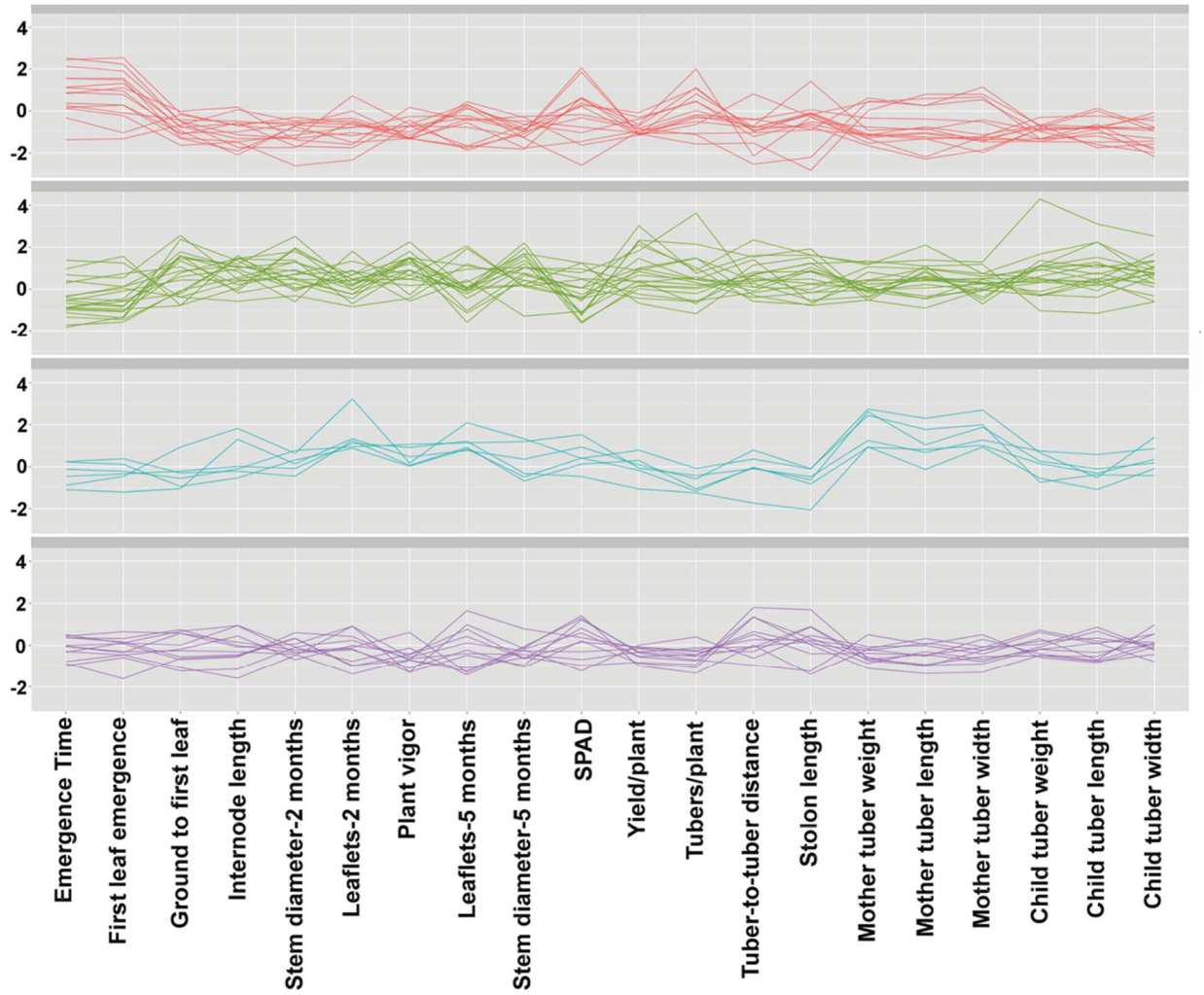

Figure 6. A multi-faceted parallel coordinate plot showing the performance of each of the genotype in a cluster, and for each trait measured on the 53 Apios genotypes at Ames, IA during 2011 and 2012. Clusters A1, A2, B1 and B2 from Figure 4 are shown in blue, green, purple and red respectively. A colored line represents the performance of each genotype for all the traits. The four clusters having distinct patterns are evident.

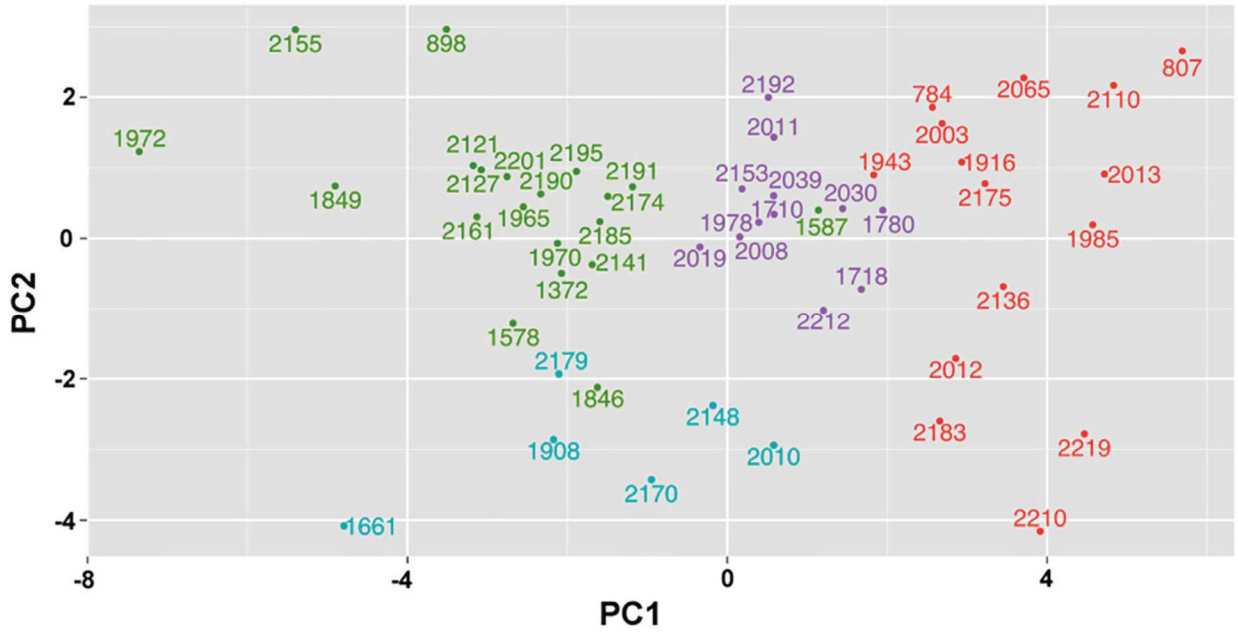

Figure 7. A plot of PC1 versus PC2 - demonstrating phenotypic diversity in the Apios collection comprising 53 genotypes, evaluated at Ames, IA during 2011 and 2012. The genotypes are colored in blue, green, purple and red representing their membership in clusters A1, A2, B1 and B2 respectively in Figure 4.

Table 1. Description of morphological and yield-related descriptors used for evaluating Apios collection at Ames, IA during 2011 and 2012, Mechanicsville, VA and Lititz, PA in 2013.

Trait class	Trait, unit	Descriptor
Emergence	Time, week	Number of weeks from planting to emergence of the vine
	First leaf, week	Number of weeks from planting to emergence of the first opened leaf
Shoot morphology	Ground to first leaf, cm	Distance from ground to the first leaf
	Internode length, cm	Average of three internode lengths (first to fourth node)
	Stem diameter - 2 and 5 months, mm	Diameter of the stem (at the ground level) measured two and five months after planting
	Plant vigor†, units	Relative vigor recorded as low (1), intermediate (2), and high (3) five months after planting
Leaf morphology and chlorophyll	Leaflets - 2 and 5 months, count	Average number of leaflets per petiole measured two and five months after planting
	Soil plant analysis development (SPAD) †, SPAD units	SPAD chlorophyll measurements recorded on the 3rd or 4th leaf (from the shoot tip) of the plant five months after planting. SPAD measurements indicate plant health through measurement of chlorophyll content of the leaves
Yield	Yield/plant†‡, g	Total mass of the below-ground portion, which includes stolons and tubers, per plant
	Tubers/plant†‡, count	Number of tubers per plant
Stolon morphology	Tuber-to-tuber distance, cm	Ratio of distance between tubers on a long stretch of stolon and the number of tubers within that distance
	Stolon length, cm	Approximate length of the stolon
Mother tuber	Weight‡, length and width, g	Weight, length and width of the mother tuber
Child tuber	Weight‡, length and width, g	Weight, length and width of the largest child tuber

†Multiple measurements made on these traits on a whole plot basis, unlike other traits, which were recorded on individual plants in the plot.

‡Traits used to evaluate the subset of the Apios collection at Mechanicsville, VA and Lititz, PA in 2013.

Table 2. REML variance component estimates[†], standard error (S.E.), and broad-sense heritability (H^2) of morphology and yield-related traits measured on the *Apios americana* collection at Ames, IA during 2011 and 2012.

Trait	R/Y	S.E.	G	S.E.	Error	S.E.	P	H^2
Emergence time, week	0.09^{ns}[‡]	0.11	0.15^{ns}	0.09	0.90	0.11	0.37	0.40
First leaf emergence, week	0.06^{ns}	0.07	0.09^{ns}	0.06	0.72	0.09	0.27	0.32
Ground to first leaf, cm	0.30^{ns}	0.52	2.66^{*}	1.13	11.10	1.30	5.43	0.49
Internode length, cm	0.14^{ns}	0.21	2.31^{**}	0.63	3.21	0.38	3.11	0.74
Stem diameter - 2 months, mm	0.03^{ns}	0.03	0.10^{**}	0.03	0.15	0.02	0.14	0.73
Leaflets - 2 months, count	0.01^{ns}	0.01	0.28^{**}	0.07	0.31	0.04	0.36	0.78
Plant vigor, units	0.01^{ns}	0.01	0.12^{**}	0.04	0.24	0.03	0.18	0.67
Leaflets - 5 months, count	0.02^{ns}	0.02	0.28^{**}	0.07	0.19	0.02	0.33	0.85
Stem diameter - 5 months, mm	0.10^{ns}	0.12	0.29^{**}	0.10	0.83	0.10	0.49	0.58
SPAD, SPAD units	0.00^{ns}	0.10	2.19^{**}	0.74	5.32	0.63	3.52	0.62
Yield/plant, g	1462.99^{ns}	1710.11	9517.60^{**}	2505.81	11705.75	1377.89	12444.04	0.76
Tubers/plant, count	4.99^{ns}	6.87	44.20^{**}	13.87	91.81	10.84	67.15	0.66
Tuber-to-tuber distance, cm	0.00^{ns}	0.01	1.62^{**}	0.49	3.23	0.38	2.43	0.67
Stolon length, cm	5.09^{ns}	24.77	200.32^{*}	98.13	1004.25	118.75	451.39	0.44
Mother tuber weight, g	554.04^{ns}	580.85	1176.35^{**}	312.15	1418.90	167.67	1531.07	0.77
Mother tuber length, cm	0.20^{ns}	0.21	0.72^{**}	0.18	0.69	0.08	0.90	0.81
Mother tuber width, cm	0.22^{ns}	0.23	0.67^{**}	0.16	0.58	0.07	0.81	0.82
Child tuber weight, g	31.33^{ns}	37.87	143.79^{**}	44.54	305.08	35.76	220.06	0.65
Child tuber length, cm	0.01^{ns}	0.04	0.60^{**}	0.17	1.02	0.12	0.85	0.70
Child tuber width, cm	0.07^{ns}	0.08	0.13^{*}	0.05	0.45	0.05	0.24	0.53

[†]R, replicate; Y, year; G, genotype; P, phenotype.
[‡]ns, not significant at the 0.05 probability level.
[*]Significant at the 0.05 probability level.
[**]Significant at the 0.01 probability level.

Table 3. REML variance component estimates[†], standard error (S.E.), and broad-sense heritability (H^2) of four yield-related traits measured on the *Apios americana* collection at Ames, IA during 2011 and 2012, Mechanicsville, VA and Lititz, PA in 2013.

Trait	R/E	S.E.	G	S.E.	GXE	S.E.	Error	S.E.	P	H^2
Yield/plant, g	1144.06 ^{ns‡}	1204.84	5337.72*	2653.03	14796.19**	3896.11	12155.69	1964.23	10556.23	0.51
Tubers/plant, count	4.08 ^{ns}	5.02	40.74**	15.80	71.19**	21.02	80.10	12.72	68.55	0.59
Mother tuber weight, g	396.58 ^{ns}	345.80	1228.42**	439.68	2281.49**	495.54	1220.98	192.86	1951.42	0.63
Child tuber weight, g	22.67 ^{ns}	23.70	72.70*	28.36	24.14 ^{ns}	31.06	268.75	34.23	112.33	0.65

[†]R, replicate; E, environment; G, genotype; P, phenotype.

[‡]ns, not significant at the 0.05 probability level.

[*]Significant at the 0.05 probability level.

[**]Significant at the 0.01 probability level.

Table 4. REML variance component estimates[†], standard error (S.E.), and broad-sense heritability (H^2) of morphology and yield-related traits measured on the *Apios americana* collection grown in 304.8 mm [12 in.] pots and 381 mm [15 in.] grow-bags at Ames, IA during 2011-2013.

Trait	R/T	S.E.	G	S.E.	GXT	S.E.	Error	S.E.	P	H^2
Emergence time, week	0.40 ^{ns‡}	0.62	0.01 ^{ns}	0.26	0.69 ^{ns}	0.57	1.89	0.46	0.83	0.01
First leaf emergence, week	0.12 ^{ns}	0.22	0.25 ^{ns}	0.27	0.53 ^{ns}	0.55	1.82	0.47	0.97	0.25
Ground to first leaf, cm	1.60 ^{ns}	2.81	9.34*	4.36	8.66 ^{ns}	6.85	19.52	5.25	18.55	0.50
Internode length, cm	0.29 ^{ns}	0.56	1.51*	0.71	0.00 ^{ns}	1.07	5.55	1.14	2.90	0.52
Stem diameter - 2 months, mm	0.18 ^{ns}	0.25	0.04 ^{ns}	0.03	0.07 ^{ns}	0.04	0.11	0.02	0.10	0.36
Leaflets - 2 months, count	0.34 ^{ns}	0.49	0.28**	0.08	0.00 ^{ns}	0.06	0.39	0.07	0.37	0.74
Plant vigor, units	0.01 ^{ns}	0.02	0.00 ^{ns}	0.02	0.03 ^{ns}	0.04	0.20	0.04	0.06	0.00
Leaflets - 5 months, count	0.08 ^{ns}	0.13	0.13**	0.05	0.00 ^{ns}	0.06	0.39	0.07	0.22	0.57
Stem diameter - 5 months, mm	0.63 ^{ns}	0.90	0.04 ^{ns}	0.05	0.14 ^{ns}	0.08	0.29	0.06	0.18	0.22
Yield/plant, g	202.98 ^{ns}	238.59	0.00 ^{ns}	150.60	180.95 ^{ns}	256.28	1826.84	256.25	547.19	0.00
Tubers/plant, count	2.65 ^{ns}	3.23	7.77 ^{ns}	4.42	7.92 ^{ns}	4.99	30.01	4.21	19.22	0.40
Tuber-to-tuber distance, cm	0.30 ^{ns}	0.44	0.00 ^{ns}	0.48	0.19 ^{ns}	0.88	6.96	0.99	1.83	0.00
Stolon length, cm	75.35 ^{ns}	88.35	11.37 ^{ns}	47.19	0.00 ^{ns}	78.76	637.94	90.92	170.86	0.07
Mother tuber weight, g	873.78 ^{ns}	883.75	213.86**	70.51	0.00 ^{ns}	59.82	515.77	72.29	342.80	0.62
Mother tuber length, cm	0.58 ^{ns}	0.59	0.33**	0.10	0.00 ^{ns}	0.08	0.69	0.10	0.50	0.65
Mother tuber width, cm	0.61 ^{ns}	0.62	0.19**	0.06	0.00 ^{ns}	0.05	0.46	0.06	0.31	0.63
Child tuber weight, g	11.83 ^{ns}	13.49	4.71 ^{ns}	4.67	0.00 ^{ns}	8.27	86.32	12.12	26.29	0.18
Child tuber length, cm	0.04 ^{ns}	0.05	0.06 ^{ns}	0.05	0.00 ^{ns}	0.08	0.84	0.12	0.27	0.23
Child tuber width, cm	0.07 ^{ns}	0.08	0.03 ^{ns}	0.03	0.00 ^{ns}	0.05	0.53	0.07	0.16	0.19

[†]R, replicate; T, treatment; G, genotype; P, phenotype.
[‡]ns, not significant at the 0.05 probability level.
[*]Significant at the 0.05 probability level.
[**]Significant at the 0.01 probability level.

Table 5. REML based estimates of LS means, ranges, standard deviation (SD), and coefficient of variation (CV) of 53 Apios genotypes for 20 descriptors evaluated in Ames, IA during 2011 and 2012.

Trait	Mean ± SEM[†]	Min.	Min. - Genotypes[‡]	Max.	Max. - Genotypes[§]	SD	CV
Emergence time, week	6.9 ± 0.0	6.5	2121, 898	7.5	1985, 2183	0.2	0.0
First leaf emergence, week	7.5 ± 0.0	7.2	2192, 898	7.9	2183, 1985	0.2	0.0
Ground to first leaf, cm	5.3 ± 0.2	3.4	807, 2065	8.1	1846, 2121	1.1	0.2
Internode length, cm	7.3 ± 0.2	4.6	2136, 2065	9.6	1661, 2174	1.3	0.2
Stem diameter - 2 months, mm	2.9 ± 0.0	2.2	807, 1985	3.6	1972, 2141	0.3	0.1
Leaflets - 2 months, count	5.9 ± 0.1	4.8	807, 2110	7.4	1661, 1372	0.5	0.1
Plant vigor, units	1.6 ± 0.0	1.2	2183	2.2	2155, 2190	0.3	0.2
Leaflets - 5 months, count	5.7 ± 0.1	4.8	2110, 2065	6.7	1661, 1578	0.5	0.1
Stem diameter - 5 months, mm	5.5 ± 0.1	4.8	807, 2175	6.4	2155, 2161	0.4	0.1
SPAD, SPAD units	29.4 ± 0.2	26.4	2012, 2136	31.8	2110, 2210	1.2	0.0
Yield/plant, g	281.0 ± 11.6	183.2	1985, 2136	537.0	1972, 2155	84.5	0.3
Tubers/plant, count	19.0 ± 0.7	10.6	2210, 2212	38.4	898, 2155	5.3	0.3
Tuber-to-tuber distance, cm	7.7 ± 0.1	5.1	2219, 807	10.1	1849, 2192	1.0	0.1
Stolon length, cm	94.0 ± 1.3	67.9	2210, 2219	111.7	2155, 2192	9.2	0.1
Mother tuber weight, g	72.7 ± 4.1	23.5	2110, 1985	154.4	1661, 2170	29.8	0.4
Mother tuber length, cm	5.8 ± 0.1	4.0	2110, 807	7.5	1661, 1849	0.8	0.1
Mother tuber width, cm	4.1 ± 0.1	2.6	1985, 2110	6.1	1661, 1908	0.7	0.2
Child tuber weight, g	37.6 ± 1.3	23.7	807, 1916	78.8	1972, 1849	9.6	0.3
Child tuber length, cm	5.0 ± 0.1	3.9	2210, 2219	7.0	1972, 2155	0.6	0.1
Child tuber width, cm	3.2 ± 0.0	2.7	2136, 2219	3.9	1972, 898	0.3	0.1

[†]SEM, standard error of mean.
[‡]Min. - Genotypes, genotypes with the lowest value for each trait.
[§]Max. - Genotypes, genotypes with the highest value for each trait.

Table 6. REML based estimates of LS means, ranges, standard deviation (SD), and coefficient of variation (CV) for four descriptors evaluated on Apios collection at Ames, IA, Mechanicsville, VA and Litiz, PA during 2011-2013.

Trait	Environment	Year	Sample size	Mean ± SEM[†]	Min.	Max.	SD	CV
Yield/plant, g	Ames, IA	2011	50	174.1 ± 12.8	62.2	453.4	90.3	0.5
Tubers/plant, count	Ames, IA	2011	50	14.6 ± 0.9	4.6	42.9	6.6	0.5
Mother tuber weight, g	Ames, IA	2011	50	40.1 ± 3.5	10.2	135.2	24.8	0.6
Child tuber weight, g	Ames, IA	2011	50	26.5 ± 1.0	15.5	57.2	7.3	0.3
Yield/plant, g	Ames, IA	2012	53	388.7 ± 14.5	236.5	681.1	105.8	0.3
Tubers/plant, count	Ames, IA	2012	53	23.2 ± 1.2	9.9	45.7	8.4	0.4
Mother tuber weight, g	Ames, IA	2012	53	106.4 ± 6.7	37.5	260.4	49.0	0.5
Child tuber weight, g	Ames, IA	2012	53	49.0 ± 1.1	35.8	77.4	7.9	0.2
Yield/plant, g	Mechanicsville, VA	2013	36	1056.8 ± 31.9	696.3	1515.4	191.2	0.2
Tubers/plant, count	Mechanicsville, VA	2013	36	55.7 ± 2.2	27.5	85.2	13.4	0.2
Mother tuber weight, g	Mechanicsville, VA	2013	36	184.3 ± 15.3	87.8	467.3	92.0	0.5
Child tuber weight, g	Mechanicsville, VA	2013	36	40.6 ± 1.1	28.4	62.0	6.4	0.2
Yield/plant, g	Lititz, PA	2013	20	91.9 ± 6.8	39.9	163.7	30.5	0.3
Tubers/plant, count	Lititz, PA	2013	20	8.6 ± 1.2	1.3	21.1	5.3	0.6
Mother tuber weight, g	Lititz, PA	2013	20	44.3 ± 4.5	5.4	75.3	19.9	0.5
Child tuber weight, g	Lititz, PA	2013	20	6.6 ± 1.2	0.0	23.7	5.2	0.8

[†]SEM, standard error of mean.

Table 7. Phenotypic correlations[†] among the four traits recorded at Ames, IA during 2011 and 2012, Mechanicsville, VA and Lititz, PA in 2013.

	Trait_ Environment[‡]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	Yield/plant_E1	NA[§]	0.00	0.00	0.00	0.00	0.09	0.01	0.00	0.04	0.71	0.12	0.00	0.02	0.17	0.68	0.00
2	Tubers/plant_ E1	0.77	NA	0.90	0.01	0.00	0.00	0.83	0.06	0.29	0.05	0.86	0.33	0.35	0.00	0.00	0.60
3	Mother tuber weight_E1	0.42	0.02	NA	0.01	0.10	0.03	0.00	0.04	0.78	0.01	0.00	0.01	0.83	0.01	0.00	0.15
4	Child tuber weight_E1	0.75	0.37	0.39	NA	0.00	0.64	0.00	0.00	0.04	0.52	0.05	0.00	0.01	0.68	0.21	0.00
5	Yield/plant_E2	0.69	0.52	0.23	0.72	NA	0.00	0.04	0.00	0.04	0.34	0.27	0.00	0.22	0.37	0.32	0.00
6	Tubers/plant_ E2	0.24	0.57	-0.30	0.07	0.49	NA	0.00	0.93	0.27	0.00	0.13	0.61	0.77	0.00	0.00	0.98
7	Mother tuber weight_E2	0.35	0.03	0.73	0.40	0.28	-0.42	NA	0.01	0.50	0.00	0.00	0.01	0.47	0.00	0.02	0.30
8	Child tuber weight_E2	0.62	0.27	0.30	0.95	0.72	0.01	0.38	NA	0.07	0.51	0.16	0.00	0.04	0.36	0.16	0.00
9	Yield/plant_E3	0.35	0.18	0.05	0.34	0.34	0.19	0.12	0.30	NA	0.00	0.57	0.01	0.01	0.58	0.84	0.45
10	Tubers/plant_ E3	0.06	0.34	-0.44	-0.11	0.16	0.69	-0.48	-0.11	0.46	NA	0.00	0.38	0.20	0.00	0.00	0.68
11	Mother tuber weight_E3	0.27	-0.03	0.63	0.33	0.19	-0.26	0.70	0.24	0.10	-0.49	NA	0.04	0.26	0.01	0.02	0.30
12	Child tuber weight_E3	0.64	0.17	0.43	0.96	0.72	-0.09	0.42	0.97	0.42	-0.15	0.34	NA	0.04	0.18	0.02	0.00
13	Yield/plant_E4	0.51	0.22	0.05	0.55	0.28	0.07	-0.17	0.46	0.61	0.33	-0.29	0.50	NA	0.24	0.18	0.03
14	Tubers/plant_ E4	0.32	0.76	-0.57	-0.10	0.21	0.81	-0.61	-0.22	0.14	0.78	-0.64	-0.34	0.27	NA	0.00	0.61
15	Mother tuber weight_E4	-0.10	-0.62	0.64	0.29	-0.24	-0.82	0.50	0.33	0.05	-0.68	0.54	0.58	0.31	-0.72	NA	0.17
16	Child tuber weight_E4	0.66	0.13	0.34	0.97	0.60	0.00	0.25	0.96	0.20	-0.11	0.27	0.95	0.49	-0.12	0.32	NA

[†]The above-diagonal elements denote *P*-values and the below diagonal elements represent correlation coefficients.

[‡]E1, Ames, IA 2011; E2, Ames, IA 2012; E3, Mechanicsville, VA; E4, Lititz, PA.

[§]Not applicable.

Table 8: Rank correlations of four traits measured at Ames, IA during 2011 and 2012, Mechanicsville, VA and Lititz, PA in 2013.

	Coefficients	*P*-values
Yield/plant - Ames, IA 2011 vs. Ames, IA 2012	0.53	0.00
Tubers/plant - Ames, IA 2011 vs. Ames, IA 2012	0.57	0.00
Mother tuber weight - Ames, IA 2011 vs. Ames, IA 2012	0.66	0.00
Child tuber weight - Ames, IA 2011 vs. Ames, IA 2012	0.80	0.00
Yield/plant - Ames, IA 2011 vs. Mechanicsville, VA	-0.12	0.39
Tubers/plant - Ames, IA 2011 vs. Mechanicsville, VA	0.04	0.80
Mother tuber weight - Ames, IA 2011 vs. Mechanicsville, VA	0.37	0.01
Child tuber weight - Ames, IA 2011 vs. Mechanicsville, VA	0.19	0.18
Yield/plant - Ames, IA 2011 vs. Lititz, PA	-0.25	0.07
Tubers/plant - Ames, IA 2011 vs. Lititz, PA	-0.10	0.49
Mother tuber weight - Ames, IA 2011 vs. Lititz, PA	-0.15	0.28
Child tuber weight - Ames, IA 2011 vs. Lititz, PA	-0.16	0.25
Yield/plant - Ames, IA 2012 vs. Mechanicsville, VA	-0.11	0.44
Tubers/plant - Ames, IA 2012 vs. Mechanicsville, VA	0.15	0.28
Mother tuber weight - Ames, IA 2012 vs. Mechanicsville, VA	0.30	0.03
Child tuber weight - Ames, IA 2012 vs. Mechanicsville, VA	0.29	0.03
Yield/plant - Ames, IA 2012 vs. Lititz, PA	-0.38	0.00
Tubers/plant - Ames, IA 2012 vs. Lititz, PA	-0.03	0.83
Mother tuber weight - Ames, IA 2012 vs. Lititz, PA	-0.26	0.06
Child tuber weight - Ames, IA 2012 vs. Lititz, PA	-0.22	0.11
Yield/plant - Mechanicsville, VA vs. Lititz, PA	0.08	0.58
Tubers/plant - Mechanicsville, VA vs. Lititz, PA	0.16	0.26
Mother tuber weight - Mechanicsville, VA vs. Lititz, PA	0.20	0.16
Child tuber weight - Mechanicsville, VA vs. Lititz, PA	0.09	0.51

Table 9. The first four principal components (PCs) of the morphological and yield-related traits measured on the Apios collection at Ames, IA during 2011 and 2012.

Trait	PC1	PC2	PC3	PC4
Emergence time, week	0.21	-0.18	-0.51†	0.12
First leaf emergence, week	0.17	-0.19	-0.56	0.16
Ground to first leaf, cm	-0.21	-0.02	0.03	-0.11
Internode length, cm	-0.27	0.01	0.19	0.01
Stem diameter - 2 months, mm	-0.30	0.02	-0.19	0.03
Leaflets - 2 months, count	-0.20	-0.31	0.12	-0.33
Plant vigor, units	-0.26	0.02	0.12	0.29
Leaflets - 5 months, count	-0.12	-0.29	0.00	-0.39
Stem diameter - 5 months, mm	-0.27	0.00	-0.08	0.08
SPAD, SPAD units	0.06	-0.08	0.33	0.17
Yield/plant, g	-0.27	0.16	-0.14	0.29
Tubers/plant, count	-0.08	0.40	0.03	0.36
Tuber-to-tuber distance, cm	-0.22	0.10	-0.26	-0.38
Stolon length, cm	-0.18	0.31	-0.13	-0.29
Mother tuber weight, g	-0.20	-0.39	0.04	0.23
Mother tuber length, cm	-0.24	-0.30	-0.13	0.11
Mother tuber width, cm	-0.19	-0.41	0.06	0.24
Child tuber weight, g	-0.27	0.08	-0.14	0.03
Child tuber length, cm	-0.26	0.20	-0.25	-0.01
Child tuber width, cm	-0.28	0.07	0.06	-0.04
Variance	8.80	3.01	1.58	1.35
Proportion of variance explained	0.44	0.15	0.08	0.07
Cumulative proportion of variance	0.44	0.59	0.67	0.74

†Traits that contribute to each PC are highlighted in gray.

CHAPTER 3. GENOMICS-ASSISTED CHARACTERIZATION OF A BREEDING COLLECTION OF *APIOS AMERICANA*, AN EDIBLE TUBEROUS LEGUME

A manuscript to be submitted to Nature Communications

Vikas Belamkar[1,2], Andrew D. Farmer[3], Nathan T. Weeks[4], Scott R. Kalberer[4], William J. Blackmon[5], Steven B. Cannon[2,4]

[1]Interdepartmental Genetics, Iowa State University, Ames, IA 50011, USA

[2]Department of Agronomy, Iowa State University, Ames, IA 50011, USA

[3]National Center for Genome Resources, Santa Fe, NM 87505, USA

[4]United States Department of Agriculture - Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

[5]5097 Studley Rd, Mechanicsville, VA 23116, USA

3.1 Abstract

For species with potential as new crops, rapid improvement may be facilitated by new genomic methods. Apios (*Apios americana* Medik.), once a staple food source of Native American Indians, produces protein-rich tubers, tolerates wide range of soils, and symbiotically fixes nitrogen. We report the first high-quality *de novo* transcriptome assembly, an expression atlas, and a set of 58,154 SNP and 39,609 gene expression markers (GEMs) for characterization of a breeding collection. Both SNPs and GEMs identify six

genotypic clusters in the collection. Transcripts mapped to the *Phaseolus vulgaris* genome – another phaseoloid legume with the same chromosome number - provide provisional genetic locations for 46,852 SNPs. Linkage disequilibrium decays rapidly within 10 kb (based on the provisional genetic locations), supporting outcrossing reproduction. SNPs and GEMs identify more than 21 marker-trait associations for at least 11 traits. These results provide an example of a holistic approach for mining plant collections.

3.2 Introduction

Many un- or semi-domesticated edible plants have valuable characteristics not found in existing major crops. Such underutilized plants include those adapted to extreme climatic conditions, unusual soil types or climates, or adaptations for resistance against a range of biotic and abiotic stresses. However, domestication of any species remains a formidable challenge.

Happily, there have been recent examples of rapid progress and success of a few underutilized crops. Ortiz-Ceballos et al.[1] showed the use of an underutilized legume cover-crop, *Mucuna pruriens* subsp. *utilis*, in rotation with maize, which improved soil fertility and increased yield of maize by nearly 60%. Quinoa (*Chenopodium quinoa*) breeding has rapidly progressed in the last 4 to 5 years, along with recent wide acceptance in the society[2, 3, 4]. Perennial grains are being explored to reduce soil degradation and water contamination simultaneously[5]. DeHaan's research at The Land Institute to domesticate perennial intermediate wheat grass (Kernza™; *Thinopyrum intermedium*) has been quite successful,

with yields increasing by ~77% with two rounds of phenotypic selection[6]. These examples show that rapid improvement of underutilized crops is feasible.

We describe application of next-generation sequence (NGS) to develop genomic resources and characterize a breeding collection of the underutilized legume *Apios americana*. Apios was a staple crop of North American Indians, and was evaluated as a tuber crop by Blackmon and Reynolds in the 1980s (ref. 7, 8). We have continued the research of Blackmon and Reynolds, and recently completed rigorous phenotypic evaluation of genotypes remaining from Blackmon and Reynolds's breeding program[9]. Apios is a perennial legume that produces a podded fruit aboveground (like other legumes), and also makes tubers belowground that are of primary interest. The tubers form on stolons (modified stems, like in potato)[9]. Some exceptional nutritional qualities of Apios tubers include high protein content (11 to 14% on dry defatted basis), long shelf life at 4 °C (>1 yr), high amounts of novel isoflavones, and low levels of reducing sugars[9].

Apios is native to the central and eastern half of North America. It grows along creeks, rivers and lakes, but can be grown on farmland. The plant is adapted to varied climatic conditions in its geographical range, and is capable of forming nodules with rhizobia, and fixing atmospheric nitrogen in the soil through symbiotic nitrogen fixation[10]. Both diploid (2n = 2x = 22) and triploid (2n = 3x = 33)[11] populations exist, but they are morphologically indistinguishable[12]. Triploids are sterile and propagate asexually via tubers, whereas diploids appear to be generally fertile and may propagate either clonally by tubers, or sexually via seeds[12]. The flowers have a complex structure and pollination is achieved with an explosive tripping mechanism[12, 13]. Bruneau and Anderson[12] observed low fruit set in diploids, and attributed it to partial self-incompatibility and the low rate of floral visits by the

apparent primary pollinators, bees in the Megachile genus. Based on their results, Bruneau and Anderson[12] suggested an outcrossing mode of reproduction in Apios.

The first effort to develop superior cultivars included collection and characterization of germplasm, followed by cycles of hybridizations and selections, by Blackmon and Reynolds during 1985-1994 at Louisiana State University Agricultural Experiment Station in Baton Rouge, LA (Fig. 1)[7, 8]. Their breeding effort has led to a collection of improved genotypes that are high yielding, with a number of favorable characteristics. We recently conducted phenotypic evaluation of the collection, in multiple years and in multiple environments and growing conditions[9]. In the present study, we have generated extensive genomic resources using RNA-Seq, and combined it with the previously generated phenotypic data to enable selection and cultivar development, to help speeding up the improvement and domestication of Apios. The specific goals of this work include: (1) Building a *de novo* reference transcriptome assembly; (2) Developing a gene expression catalog; (3) Identifying SNPs, gene expression markers (GEMs), and genotyping the collection using RNA-Seq; (4) Investigating heterozygosity, pedigree and population structure in the collection; (5) Examining linkage disequilibrium in the collection; (6) Identifying marker-trait associations using SNPs and gene expression markers; and (7) Combining phenotypic and genetic data to make parental selections for subsequent cultivar development. The results obtained in this study will have broader implications for plants with limited genomic resources, as well as for staple crops whose collections can be further mined for crop diversity and cultivar development.

3.3 Results

3.3.1 Development and evaluation of an Apios breeding collection. The plant material utilized in this study comprised 52 genotypes from Blackmon and Reynolds's breeding program conducted at Louisiana State University Agricultural Experiment Station in Baton Rouge, LA during 1985-1994 (Fig. 1). Field books of the breeding program described collection of germplasm, pollinations performed, and phenotypic selections of ~20,000 plants (Fig. 1a). Of those evaluations, 53 genotypes remained as of 2010 (of which 52 were carried forward in the present study). Partial pedigree information was traceable for 35 of the 53 genotypes in the collection (Fig. 1b). "Good-performing" genotypes were repeatedly favored during breeding and this resulted in a number of primary founder genotypes. For instance, line "034" contributed to 22 of the 52 genotypes, and "006" to 8 of the 52 genotypes (Fig. 1b). Line 034 performed consistently under different growing conditions, and therefore was selected for release in the late 1980s. All of the genotypes in the collection were derived through pollinations, and thus are likely to be diploids. Twenty-five of the 53 genotypes tested using flow cytomtery had an average genome size of 1644 ± 34 Mb, whereas wild accessions collected from Iowa, New York, and Quebec in Canada had an average genome size of 2380 ± 28 Mb (Supplementary Table S1). The genome size of the genotypes in the Blackmon and Reynolds collection is nearly two-third the average genome size of the wild accessions, suggesting our particular wild accessions to be triploids, and the Blackmon and Reynolds genotypes to be diploids. The diploid and triploid genome size estimates are comparable to the results previously published[14]. Recent field evaluation of the 53 genotypes from the Blackmon and Reynolds collection found high amounts of phenotypic diversity for

18 of the 20 traits measured[9]. The REML-based LS means generated for each trait across both years in that phenotyping study[9] are used in this work.

3.3.2 *De novo* transcriptome assembly, annotation and expression catalog. A transcriptome assembly for *A. americana* was built by sequencing 14 above- and belowground samples, from six tissues, from line 2127 (Supplementary Data S1). Sequencing of the 14 samples on Illumina GAIIx/HiSeq generated ~210 million reads, which were then assembled into 96,560 transcripts, and 48,615 components (generally corresponding with genes) (Table 1). The length of the transcripts ranged from 201 to 15,850 bp, with an average of 1,173 bp, and N50 of 1,863 bp (Supplementary Fig. S1). The statistics obtained for this assembly are at least on par with most of the *de novo* transcriptome assemblies published for other plant species to date[15, 16, 17, 18, 19]. Transcript analysis indicates that more than 14% of the transcripts in the assembly are nearly full-length (>80% alignment coverage) as compared to soybean and common bean peptides (14,905 (15.4%) and 14,084 (14.6%), respectively). The percentage of transcripts matching other legumes with genome sequences and annotations - or Arabidopsis - ranged from 61.1% to 67.5% (Supplementary Fig. S2, Supplementary Data S2). The percentage of Apios transcripts matching *P. vulgaris* proteins is 66.5%, while the percentage of *P. vulagris* proteins matching Apios transcripts is 66% (Supplementary Fig. S2). Such a high one-to-one match between Apios transcripts and common bean peptides suggests relative completeness of the Apios transcriptome assembly. Extraction of the putative coding regions from the transcripts generated 23,691 unique peptides (Supplementary Data S3), of which greater than 70% were assigned a function using Swiss-Prot and Pfam databases, >57% using eggNOG and Gene Ontology. Nearly 24%

contained a transmembrane helix, and ~9% were tentatively marked as signal peptides (Table 1; Supplementary Data S4).

We also constructed a gene expression catalog using an average of 14.6 million reads from each of the 11 samples, including six tissues, each with two biological replicates except for flower, which had one replicate (Supplementary Table S2). Nearly 90% of the quality-trimmed reads from each of these samples could be mapped back to the assembly (Supplementary Table S2). We found 56,735 transcripts (Supplementary Data S5), and 28,738 components (Supplementary Data S6) expressed in at least one of the 11 samples. A heat map utilizing expression of 1,000 transcripts with highest variances clustered the samples into two primary subgroups, one consisting of the aboveground tissues, and the other comprising the belowground tissues (Fig. 2).

3.3.3 Marker discovery, validation and genotyping of the collection. The 52 genotypes in the collection were grown for three months under field conditions. Their leaf transcriptomes were sequenced on Illumina Hi-Seq, and 1.32 billion reads, with an average of 25.3 million reads per genotype (Table 2, Supplementary Data S1). Additionally, transcriptome sequencing of the pooled (shoot and root tissue) sample from four genotypes (as biological replicates), grown for 1.5 months under greenhouse conditions, generated 45.71 million reads. On an average, >88.6% of the reads from each of the 56 samples were mapped to the *de novo* transcriptome assembly (Supplementary Data S7). Using stringent filtering criteria (described in the Methods), 58,154 high-quality SNP markers were identified in the collection, in 9,338 components (Table 2). The variant files generated at each filtering stage are provided as Variant Call Format (VCF) files (Supplementary Data S8 - S10). The average reproducibility of the 58,154 high-quality SNP markers tested using four biological replicates

was 90.6% (Table 2; Supplementary Table S3). This value is comparable to the reproducibility (78% to 92.9 %) observed in other recent studies[20, 21, 22, 23, 24]. The allele frequency of the G and C alleles was slightly higher in the dataset than the A and T alleles, which is expected, considering the fact that these SNP markers are derived from the transcribed regions of the genome (Supplementary Table S4). The SNP marker and the genotype summary indicated the average missing percentage per accession as 1.5%, and the average heterozygosity per accession as 37.6 (Supplementary Data S11 - S12). Lastly, the reads from each of the 52 genotypes mapped to the reference assembly provided 39,609 transcripts that were expressed in at least one genotype, and are used as gene-expression-markers (GEMs) for performing structure and association analyses (Supplementary Data S13 - S14).

3.3.4 Diversity, inbreeding and pedigree in the collection. Two commonly used measures of diversity were estimated: (1) the average pair-wise divergence among genotypes, or nucleotide diversity per bp, π (pi), was 0.35; and (2) the expected number of polymorphic sites per nucleotide, or expectation of π, θ (theta), was 0.22. These diversity estimates are somewhat higher than that of four soybean populations (including elite North American soybean cultivars, Asian landrace founders of these elite cultivars, Asian landraces (with no known relationship to the landrace founders), and accessions of the wild progenitor species *Glycine soja*[25]), further corroborating high nucleotide diversity in the collection. In addition, Tajima's D was 2.16. Tajima's D is a normalized measure of the difference between observed (π) and expected (θ) nucleotide diversity. A large positive value of this measurement suggests either that selection has maintained variation in the population, or that

the population has contracted – either of which may have occurred during development of this breeding collection.

Inbreeding coefficient estimates for each line are analogous to the proportion of heterozygous loci in each line, but in the opposite direction – that is, the correlation coefficient is -1.0 between the inbreeding coefficients and heterozygosity of each line (Supplementary Data 12). The inbreeding coefficients estimated for each genotype ranged from -0.22 to 0.15, with an average of -0.07 – which indicates the generally high heterozygosity of genotypes in the collection (Supplementary Data 12). In fact, inbreeding coefficient values of only five of the genotypes in the collection (1718, 1846, 2148, 2153 and 2170) were positive, indicating that these genotypes are more homozygous than the average of the population[26]. We noticed heterozygosity had a strong effect on yield-related traits, and higher heterozygosity corresponded with higher tubers per plant. Statistically significant (P<0.05) negative correlation was observed between inbreeding coefficients and tubers/plant (Supplementary Table S5; Supplementary Fig. S3). Positive correlations were observed between inbreeding coefficients and three phenotypic traits: leaflets measured 2 months after planting, mother tuber weight, and mother tuber width.

Using estimates of Identity-by-Descent (IBD) and the proportion of IBD values, we identified parent-child or half-sib relationships between 24 pairs of genotypes in the collection (Supplementary Table S6). Of the 24 pairs, nine pairs were validated using the partial pedigree information available for genotypes in the collection (Fig. 1b); one pair (807 and 2003) had conflicting results between IBD analysis and known pedigree information; and the remaining 14 pairs identified by IBD analysis did not have prior pedigree information. All of the pairs identified in the IBD analysis can be validated using fastSTRUCTURE

results (described in the next section) either by inspecting the proportion of genomes shared, or the clustering of genotypes in the same group.

3.3.5 Population structure of the collection. The structure of the collection was analyzed using five different approaches (see Materials and Methods for details), and at least six clusters were identified in the collection (Fig. 3). Bayesian analysis implemented using the program fastSTRUCTURE suggested the presence of five (k=5) to seven (k=7) clusters in the collection (Fig. 3a). The main difference between K=5 and K=6, is the inclusion of an additional cluster comprising of genotypes 784, 2012, 2019 and 2219. Genotype 784 is the maternal parent of the genotypes 2012 and 2019 (Fig. 1b), and thus these genotypes are expected to cluster in the same group. Hence, the additional cluster in K=6 is reasonable. In the case of K=7, the additional cluster is formed by splitting the fifth cluster of K=6. The fifth cluster in K=6 comprises of individuals that are admixed, and these further split into an additional cluster in K=7. However, admixed individuals have >60% proportion of membership in K=6, and there is insufficient evidence to classify these individuals as a separate cluster. Hence, there are at least six clusters in the collection based on the variational Bayesian analysis conducted using fastSTRUCTURE. By implementing the maximum likelihood approach, fastSTRUCTURE clusters 2, 3, 4, and 6 remained intact, whereas the admixed individuals from clusters 1 and 5 split up and re-grouped with the other clusters (Fig. 3b). Similar results were observed in the IBS and Ward's clustering approach with the exception that the fifth cluster split into exactly two clusters as opposed to three in maximum likelihood approach (Supplementary Fig. S4). The PCA identified six clusters, but a seventh cluster was formed which comprised of individuals from clusters 1, 2 and 5 (Supplementary Fig. S5). Interestingly, GEMs also identified structure in the collection (Fig. 4). The four

clusters (2, 3, 4, and 6) remained intact, and the same was true for cluster 1 that contained a few admixed individuals. However, in the GEM-based analysis, cluster 5 is split up and individuals re-group with other clusters. In summary, the SNP and GEM markers identify at least six clusters - of which five remain intact in fastSTRUCTURE and GEM-based phylogenies. Four of the clusters are consistent across all the five approaches. As a final step, we compared the performance of each of the five approaches in accurately clustering the genotypes based on the pedigree, and the highest success was obtained using fastSTURCTURE (Supplementary Table S7).

3.3.6 Linkage disequilibrium (LD) in the collection. Linkage disequilibrium was investigated in the collection by mapping the Apios transcripts to *P. vulgaris* chromosomes, under the assumption that the Apios and *P. vulgaris* chromosomes are generally syntenic – an assumption that we believe is warranted considering that most species in the Milletieae tribe have 11 chromosomes[27, 28], and various divergent species in the Phaseoleae are strongly collinear[29, 30]. We were able to place 46,852 of the 58,154 Apios SNP markers. These SNPs are distributed along each of the 11 chromosomes, with a slight enrichment toward chromosome ends (Fig. 5a). The enrichment at the chromosomal ends is undoubtedly because the SNPs are derived from transcripts, and the chromosome ends are gene rich[31]. The number of SNPs per chromosome ranged from 2,452 (chromosome 10) to 6,074 (chromosome 2) with an average of 4,259 (Supplementary Fig. S6). The decay of LD was investigated at different r^2 thresholds, as well as along each of the chromosome, across the genome, and along the transcripts (Fig. 5b; Supplementary Table S8). On average, LD extends up to 10 to 15 kb (at $r^2 \leq 0.15$) across the genome. Although LD decays rapidly across the genome, it is well known that the extent of LD varies in different regions of the

genome. Using a sliding window of 2 Mb along each chromosome, we identified 4,222 haplotype blocks across the genome. The average size of each haplotype block was 7.4 kb. The distribution of haplotype blocks along each of the chromosome indicated enrichment of large haplotype blocks in the pericentromeric regions (Fig. 5c). The two largest haplotype blocks (~1,500 kb) were detected in the pericentromeric region on chromosome 5 and 10. The pericentromeric regions experience lower rates of recombinations, and thus are likely to retain large-sized halpotype blocks. In summary, mapping of Apios SNPs to the *P. vulgaris* chromosomes provided putative location information for ~81% of the SNPs, and using these SNPs LD was found to decay within 10-15 kb.

3.3.7 Marker-trait associations in the collection. The association analysis was performed using five different models (see Materials and Methods), and in two different packages. We inspected the performance of each of the model by using quantile-quantile (QQ) plots. Based on the QQ plots, the most successful model (performed in the GCTA package) accounted for population structure (thereby reducing false positives), using a familial relatedness matrix in the mixed model analysis (Supplementary Fig. S7). For 19 of the 20 traits it was apparent that incorporating the familial relatedness matrix in the mixed model analysis (performed in the GCTA package) performed the best, and for one of the traits (child tuber weight), two additional models (performed in TASSEL) performed equally well. These models accounted either for subpopulations and familial relatedness together, or just familial relatedness (Supplementary Fig. S7). Thus, for marker-trait associations, we proceeded with the mixed model analysis in the GCTA package, incorporating the familial relatedness for controlling population structure. This method identifies twenty-one SNP markers to be associated with 14 phenotypic traits, including six aboveground and eight belowground traits (Table 3). Each

of the SNP markers associated with a trait originated from a different transcript. Additional details regarding favorable allele, allele frequency, effect size, and annotation of the transcript are provided in Table 3. The marker-trait associations are also displayed in Manhattan plots (Supplementary Fig. S8). Additionally, we evaluated potential parental selections for the purpose of further cultivar improvement. We identified accessions that contained beneficial alleles with large effect sizes for yield-related traits (Supplementary Table S9). Many of these accessions are among the top 10% (based on the phenotypic data) of the performers in field evaluations[9].

3.3.8 Marker-trait associations using gene expression markers (GEMs). The reads generated by sequencing the leaf transcriptome from each of the 52 genotypes in the collection were mapped to 92,092 transcripts of the 96,560 transcripts in the *de novo* reference assembly, and a normalized expression dataset was generated (Supplementary Data S13). The normalized expression dataset was further filtered for transcripts expressed in at least one genotype in the collection, which yielded 39,609 transcripts, or GEMs across the collection (Supplementary Data S14). The regression analysis performed using GEMs, and filtered by applying Bonferroni correction (adj-$P < 0.0000013$), resulted in 34 GEM-trait associations (Supplementary Fig. S9-S10). Six of these GEM-trait associations were excluded because they violated the linearity assumption of linear regression analysis (Supplementary Fig. S9-S10). Finally, 28 GEM-trait associations were identified for four aboveground, and five belowground traits (Table 4, Supplementary Fig. S11-S12). Nine GEM-trait associations (Fig. 6) are particularly interesting, and can be broadly classified into three categories - (1) The expression of isoforms of the same gene are correlated with the same trait, but one of the isoforms is positively correlated, whereas the other is negatively

correlated with the same trait. This may suggest an autoregulatory feedback mechanism, or opposite roles in controlling the phenotype by these isoforms; (2) Transcripts are expressed only in the five highest-performing genotypes (with an exception of one outlier), whereas the rest of the genotypes shown no expression, suggesting their possible role in the better performance of the genotypes for the respective trait; and (3) A transcript has lower expression (~FPKM<20) in genotypes that produce shorter child tubers, and has higher expression (~FPKM>40) in genotypes that produce longer child tubers. This indicates the probable role of the transcript in regulating the length of the child tubers. Overall, association analysis conducted using GEMs has revealed several interesting candidate genes for both above and belowground traits.

3.4 Discussion

Mining the diversity of plant collections is critical for developing cultivars. Flow cytometry analysis shows that at least 25 genotypes of the Blackmon and Reynolds collection are diploids (and we reason that all of the genotypes are diploid), while six of our wild-collected accessions from central and eastern USA, and Canada are triploids. Triploid genotypes can act as pollen donors, but do not set seed[12]. Thus, it is likely that the diploid genotypes were unintentionally selected during the Blackmon and Reynolds breeding effort. We believe the Blackmon and Reynolds collection used in this study is the largest existing breeding collection. These genotypes are a result of ~10 years of breeding, and we have observed tuber yields of more than 1,500 g per plant in the elite genotypes[9]. Significant nucleotide diversity is present in the collection based on diversity estimates, π and θ,

generated using the 58,154 high-quality SNPs produced by sequencing the leaf transcriptomes of 52 genotypes in the collection. Thus, the Blackmon and Reynolds collection has potential for continued improvement and cultivar development.

Many of tuber crops, such as potato, cassava, sweet potato, and yams, are clonally propagated. They are often highly heterozygous and exhibit hybrid vigor[32]. An advantage of clonal propagation is that once a superior hybrid is identified, it can be fixed and propagated in the heterotic state. The downside is that high heterozygosity makes these crops vulnerable to inbreeding depression, since self-pollination will generally increase the proportion of homozygous (and presumably often deleterious) alleles. In this study, a strong negative correlation was observed between "inbreeding coefficients estimated for each genotype" and tubers produced per plant; and positive correlations were observed between inbreeding coefficients and leaflets measured 2 months after planting, mother tuber weight, and mother tuber width. This suggests higher heterozygosity is correlated with larger number of tubers per plant, and lower heterozygosity is associated with traits that result in genotypes, which produce one large mother tuber/seed tuber and a few child tubers. Leaflets recorded 2 months after planting is highly correlated with mother tuber weight and length, and higher-values of these three phenotypic traits have previously been shown to be correlated with a "stout" tuber phenotype – i.e. one large mother tuber and just a few or no child tubers[9]. Hence, it is likely that Apios may exhibit hybrid vigor, and can be susceptible to inbreeding depression when forced to self-pollinate or crosses made between genetically related genotypes.

Understanding the population structure is a prime requirement for effectively utilizing genotypes from the collection for breeding purposes. Population structure needs to be accounted for in the association analysis to control for spurious associations. We find clear

population structure in the collection, with approximately six distinct genotypic clusters. Four of the clusters were consistent across all the six approaches studied. The fifth and sixth cluster identified using the program fastSTRUCTURE contained admixed individuals, and the placement of these admixed individuals was ambiguous in the other three methods (maximum likelihood, identity-by-state/Wards, PCA) utilizing the SNP dataset. Interestingly, classification on the basis of gene expression (using GEMs) was consistent with classification based on SNP genotyping. Five of the six groups identified by fastSTRUCTURE remained intact in the phylogeny generated using GEMs. This indicates that gene expression data can be effectively used as molecular markers for understanding the population structure, and its accuracy may be comparable with other widely-used SNP-based methods. Gene expression markers may be an important, underutilized resource for other species, including well-studied crops.

A clear understanding of pollination biology is essential for performing hybridization experiments. According to Bruneau and Anderson[12], Apios flowers are predominantly out-crossing. Although self-pollination may occur when the plants are made to self, the success rate is quite low. Based on the Index of Self-Incompatibility (ISI), the authors suggest existence of partial self-incompatibility with characteristics of a gametophytic self-incompatibility system. We find LD in the Apios collection generally decaying within 10 to 15 kb. Linkage disequilibrium extending to such small distances is mainly observed in cross-pollinating species[33, 34]. In the self-pollinating species, the effective recombination rate is severely reduced, and therefore the LD tends to be more extensive. Hence, the LD results obtained in this study suggest out-crossing biology, consistent with observations about the pollination by Bruneau and Anderson[12] (1988).

Marker-traits associations identified using association analysis can be utilized in at least two different ways: (1) Implementation of marker-assisted selection (MAS) in a breeding program; (2) Identification of potential parents containing favorable alleles for developing cultivars (parental selections). The high heterozygosity, and clear population structure in our dataset prompted us to test different models and packages for association analysis. Of the five models tested, the one that best accounted for population structure using "familial relatedness" was in the GCTA package. This also appeared to give the most reliable result, based on QQ-plots. Controlling for population structure by incorporating "familial relatedness" has recently been shown to be helpful in association analysis involving highly structured populations with admixture[35]. One of the limitations of our study in performing association analysis is the relatively small population size (52 individuals). Although a similar population size (53 individuals) has been used in a prior study involving *Brassica napus*, and the marker-trait association identified and validated[36], marker-trait associations identified in this study should be considered cautiously, and will require validation in another collection or in bi-parental populations grown in multiple environments.

SNP markers used in this study are from the transcribed regions, and this allows for preliminary verification of marker-trait associations using the annotations of the transcripts containing the SNPs associated with a trait. A few of the marker-trait associations that can be validated using annotations of the transcripts are: (1) S_23737486 (SNP)/comp54771_c0_seq1 (transcript; a serine/threonine protein kinase), associated with "weeks to first leaf emergence." This enzyme has been previously shown to regulate seedling germination[37, 38]. (2) S_11450514/ comp50326_c0_seq1 (a START (StAR-related lipid transfer) protein), associated with leaflets recorded 2 months after planting. START is a

lipid-binding domain and is mainly enriched in the HD-Zip genes in plants. Mutations in the two HD-Zip proteins containing START domain, *PHABULOSA* and *PHAVOLUTA* have been shown to perturb adaxial/abaxial (upper/lower) axis formation in the leaf[39]. (3) S_23095079/comp54598_c3_seq1 (Pectinesterase, a widely studied enzyme in different fruits, and a key enzyme in potato involved in regulating firmness of the tubers[40, 41]) associated with mother tuber width. (4) S_18970167/comp53339_c0_seq2 (a glucose-6-phosphate transmembrane transporter that catalyses the transfer of glucose-6-phosphate from one side of the membrane to the other), associated with mother tuber width. There are at least two hypotheses from the studies conducted in potato, which links glucose-6-phosphate transmembrane transporter to tuber size. Firstly, sucrose is a major form of sugar during transport to tubers; unloading and subsequent mobilization of sucrose during tuber initiation and enlargement involves conversion of sucrose to fructose-6-phosphate through glucose-6-phosphate as an intermediate molecule[42]. Increased sucrose mobilization in the cytosol has been shown to increase tuber number, and reduce tuber size. On the other hand, a rise in sucrose mobilization in the apoplast increases tuber size and decreases tuber number[43]. Secondly, in growing potato tubers, the oxygen levels decrease and hypoxia conditions arise, which further acts as an adaptive mechanism to conserve energy. In a recent study, Licausi et al.[44] showed that the expression of three hypoxia-responsive ethylene-responsive factor genes is associated with the drop of oxygen levels during tuber growth. The expression levels of two of the three hypoxia-related genes is highly correlated ($r > 0.9$) with two distinct glucose-6-phosphate transmembrane transporters suggesting the role of the latter in tuber growth. (5) S_30406303/comp56304_c4_seq2 (a putative transcription factor with Zinger finger domain (CCCH-type)), associated with child tuber length. The expression of these transcription

factors is highly correlated with the hypoxia-related genes described previously[44], which further suggests the involvement in tuberization.

Selection of parents is a fist key step in developing new cultivars. In our previous study[9], we had recommended potential candidate parents and crossing schemes based only on phenotypic evaluations. The first scheme was suggested for development of high-yielding genotypes, and involved the utilization of the top 10% of high-yielding genotypes to make a "good x good" cross. The genotypes that comprised the top 10% in each of the four environments include 1972, 2191, 898, 2127, 1849, 2155, 2201, 1970, and 2065 (ref. 9). Using the population structure results from this study, we can now identify genotypes that are most genetically diverse, reducing the likelihood of inbreeding depression, and promoting hybrid vigor. Based on these results, we recommend hybridizations between the genotypes in genetically distinct clusters: cluster 1 - 2191; cluster 2: 1972, 1849, and 2155; cluster 3 - 1970; and cluster 4 - 898, 2127, 2201, and 2065. In addition, favorable alleles for yield-related traits are identified in the following accessions that are among the top 10% based on their phenotype (Supplementary Table S9): Genotype 2191 for tubers/plant, 1972 (yield/plant, mother tuber length, child tuber weight and length), 1849 (yield/plant, mother tuber length), 2155 (yield/plant, tubers/plant, mother tuber length), 1970 (mother tuber length), and 898 (tubers/plant). Therefore, a cross of 1972 x 2191 or 1972 x 898 has a higher probability of producing high-yielding genotypes with a reasonably good number of large sized child tubers. A similar strategy of integrating results from population structure and association analysis should be considered for other crossing schemes, and extrapolated to the selections derived in Belamkar et al.[9], which was primarily based on the phenotypic data.

Gene expression in the genotypes can be correlated with their phenotypes, and is valuable in identification of candidate genes for phenotypic traits[36]. Recent evidence also suggests that expression data can be more powerful than SNP markers for performing predictions in smaller populations[45], and for complex traits with low heritability[46]. In this study, 28 GEM-trait associations were identified, and they include transcripts for four aboveground and five belowground traits. Interestingly, only five of these transcripts contained SNPs. Hence, these associations are valuable and would not have been captured in the association analysis conducted using SNP markers. Of the marker-trait associations identified, there were two pairs of isoforms (comp52098_c0_seq1/ comp52098_c0_seq2; comp45738_c0_seq1/ comp45738_c0_seq2), whose expression was strongly correlated with a phenotypic trait, but one of them was positively correlated and the other one was negatively correlated with the same trait. This may suggest autoregulatory feedback mechanism, or opposite roles in controlling the phenotypes by these transcripts. The expression of the transcript "comp57351_c3_seq6" was lower in genotypes with shorter child tubers, and was extremely high in genotypes with longer child tubers, resulting in clustering of the population into two groups based on its expression. This transcript is annotated as being involved in Jasmonic Acid (JA) signaling pathway, and JA's role has been well established as an effective inducer of tuberization in potato[47, 48]. Two of the transcripts (comp57301_c0_seq1 and comp55913_c0_seq2) associated with yield/plant and child tuber length were only expressed in the top five performing genotypes (1849, 1972, 2127, 2155 and 2201), including one outlier (1978). These two transcripts were mapped to Chromosome 8 of *P. vulgaris*, and are separated by 319.4 kb. Although there are no SNPs within each of these two transcripts (comp57301_c0_seq1 and comp55913_c0_seq2), there are 96 SNPs identified between them.

The LD among these 96 SNPs is relatively high, with average $r^2 = 0.17$, which is higher than the background LD value of 0.15 (Supplementary Fig. S13). The LD observed is probably not high enough to suggest selection during the Blackmon and Reynolds's breeding program, but further investigation is required to confirm this observation. Overall, GEM-based association analysis provided several interesting candidate genes, which can further provide a lead to initiate functional characterization studies for traits of interest.

In summary, we have shown combining high-throughput genomics with the phenotypic information can accelerate mining of collection for domestication of a potential new crop. We have built large amounts of genomics resources for Apios, including a high-quality reference *de novo* transcriptome assembly, an expression atlas of six tissues, and 58,154 highly reproducible SNPs and 39,609 GEMs across the collection. Both SNPs and GEMs successfully identified pedigree and population structure of the collection, and association analysis identified favorable alleles and potential candidate genes for both aboveground and belowground traits. Lastly, the success of GEMs is exciting and suggests broad utility for both new and well-studied crops.

3.5 Methods

3.5.1 Historical and morphological evaluation of the *Apios americana* collection. The collection used in the study was developed by Blackmon and Reynolds[7, 8] at Louisiana State University Agricultural Experiment Station in Baton Rouge, LA during 1985-1994 (Fig. 1a). The breeding program involved (i) collection of wild germplasm from different states of USA and Canada with majority of them originating from Louisiana; (ii) germplasm

evaluation for traits of interest; (iii) open pollinations; and (iv) phenotype-based selections for superior genotypes (Fig. 1a). The open pollinations made it impossible to determine the paternal parent, thus the genotypes growing adjacent to the female genotypes (the probable pollen donors), were generally considered to be likely paternal parents. Genotypes that produced seed were documented as female parents. Partial pedigree information derived from maternal lineage is available for 35 of the 53 accessions in the collection (Fig. 1b). The collection was recently screened[9] for phenotypic variation over multiple years, in multiple environments, and in three different growing conditions. Twenty phenotypic traits that included 10 aboveground, and 10 belowground traits were recorded for the entire collection. The REML-based least square (LS) means generated for the 20 traits recorded at Ames, IA during 2011 and 2012 were used in this study[9]. Genome size of 25 of the 53 genotypes in the collection, and six additional accessions collected in the wild from Iowa, New York and Canada was estimated using flow cytometry at the Iowa State University Flow Cytometry Facility (http://www.biotech.iastate.edu/facilities/flow/).

3.5.2 *De novo* transcriptome assembly, annotation and expression catalog. Total RNA was isolated from 14 samples including six different tissues (leaf, shoot, flower, root, mother tuber, and child tuber) from accession 2127 (Supplementary Data S1) using Qiagen RNeasy® Plant mini kit and following the manufacturer's protocol. The RNA samples were treated with Ambion® TURBO DNA-free™ DNase to get rid of any DNA contamination, and the quality and quantity were then inspected using Agilent 2100 Bioanalyzer. Illumina® libraries were prepared, and the samples were sequenced on Illumina® GAIIx, and Illumina® HiSeq 2000 platform (Supplementary Data S1). Both single- and paired-end reads of length 50 to 90 bp were generated. The reads were trimmed to exclude the Q2 (read segment control

indicator) bases from the ends of the reads using a custom script. The reads that passed the quality trimming and were at least 25 bp in length were utilized to build the *de novo* transcriptome assembly using the Trinity package[49] (release 2013-02-25). Assembled transcripts were examined for full-length transcripts, and sequence conservation across species, by performing a BLASTX search with a threshold of 1E-05 against the proteomes of six sequenced legumes, *Glycine max* (assembly v1.01, JGI Glyma 1.1 annotation), *Medicago truncatula* (v 3.5.1), *Cajanus cajan* (v 1.0), *Phaseolus vulgaris* (v 1.0), *Lotus japonicus* (v 2.5), *Cicer arietinum* (v 1.0); and the model plant *Arabidopsis thaliana* (v 10.0; release 04/16/2012); and Swiss-Prot non-redundant database (release-2013_04). Functional annotation of transcripts was performed using Trinotate[49], an annotation suite within the Trinity package. The likely coding region (ORF) in the transcripts were extracted using TransDecoder with default settings, and the option to utilize pfam using hmmscan (--search_pfam), and the minimum peptide length changed to 67 amino acids to match with the minimum length of transcripts. The functional annotation of these peptides was performed as follows: a homology search with BLASTP against the Swiss-Prot non-redundant database; protein domain identification using HMMER and Pfam-A (v 27.0); signal peptide identification using SignalP (v 4.0); transmembrane region prediction using tmHMM; and annotation by comparison with eggNog (evolutionary genealogy of genes) and Gene Ontology database. Lastly, the quality-trimmed reads from 11 samples - six different tissues and two biological replicates per tissue (except for flower sample) - were aligned to the *de novo* transcriptome assembly using Bowtie, and the abundance estimation per transcript and component (loosely termed as gene) were estimated using RNA-Seq by Expectation Maximization (RSEM)[50].

3.5.3 Genotyping of the collection using RNA-Seq. The collection was grown in 2012 at Ames, IA as described in Belamkar et al.[9]. Leaf tissue from 52 genotypes was collected 3 months after planting (on a single day). The leaf tissue from four plants of each genotype was pooled and frozen in liquid nitrogen. Total RNA was isolated from the leaf tissue and the quality was inspected as described earlier for *de novo* transcriptome assembly construction. Illumina® libraries were prepared at the DNA Facility at Iowa State University (http://www.dna.iastate.edu), and the samples were sequenced on Illumina® HiSeq at National Center for Genome Resources (NCGR), Santa Fe, NM to generate single-end, 50 bp reads (Supplementary Data S1). RNA was isolated and sequenced from pooled samples of shoots and roots of four genotypes belonging to the collection that were grown in the greenhouse for quality control purposes (Supplementary Data S1). The reads from the 56 samples were mapped to the *de novo* transcriptome assembly to identify "variants" and "transcript abundances for each genotype" using the Alpheus™ pipeline[51]. For calling single nucleotide polymorphisms (SNPs) in each genotype, a minimum read depth of ≥ 5 reads, frequency of variant allele ≥ 20 %, and average quality of bases calling the variant allele ≥ 10 was used. We further filtered SNP markers with minor allele frequency ≤ 0.1 % and maximum missing percentage ≥ 10%, and generated 58,154 high-quality SNP markers across the collection. Allele frequencies, and summary statistics for each SNP (major and minor allele frequency, missing %, heterozygous accessions %) and genotype (number of SNPs, missing SNPs %, heterozygous SNP marker sites %) were generated using the "Genotype summary" option in TASSEL v5.0 (ref. 52). The "transcript abundance per sample" was also utilized as a marker, and will be referred to as gene expression marker (GEM). A transcript

was considered expressed if the normalized expression value was greater than or equal to 2 in at least one of the genotypes, which resulted in 39,609 GEMs across the collection.

3.5.4 Diversity estimates, inbreeding and pedigree in the collection. The two commonly used diversity estimates, average pairwise divergence or nucleotide diversity per bp, π (pi), and number of segregating sites per nucleotide, θ (theta) were generated using the option "Diversity" in TASSEL. Similarly, the Tajima's D was produced in TASSEL to understand the evolutionary history of the breeding collection. The extent of heterozygosity in the collection, and its effect on phenotype was investigated by estimating inbreeding coefficients for each genotype in the collection using the "--het" option in PLINK. The correlations between inbreeding coefficients, and REML-based LS means of phenotypic traits recorded in Ames, IA in 2011-2012 was performed in R (http://www.r-project.org/). Lastly, pedigree relationships such as "parent-child" and "half-sib" relationships between genotypes in the collection were identified using estimates of identity-by-descent (IBD) as described in Stevens et al.[53]. The IBD and the proportion of IBD values were generated using the "--genome" option in PLINK with an assumption of homogenous population. The SNP data set used for IBD analysis was restricted to retain only the SNPs that were reproducible in the four control genotypes, and had minor allele frequency $\geq$ 0.1 % and maximum missing percentage $\leq$ 10%, which provided 14,321 SNP markers across the collection.

3.5.5 Population structure of the collection. Population structure of the collection was investigated using both SNPs and GEMs. Four approaches were tested using SNPs - (1) Phylogeny reconstruction using the package SNPhylo[54]. Briefly, in this package the SNP information of each genotype is used to generate sequences. The sequences are then aligned using MUSCLE, and the phylogeny tree is built using maximum likelihood. SNPhylo was

utilized with the following settings (-l 0.0, -b and -A), and the resulting tree was visualized with midpoint rooting using FigTree v1.4.2; (2) Identity-by-state (IBS) - a distance matrix based on (1-IBS) values was generated using PLINK v1.90b2pNL, and hierarchical clustering was performed using the Ward's linkage in R; (3) Principal component analysis (PCA) - performed in the program GCTA v1.24 (ref. 55), and the plot of PC1 versus PC2 was made in R; (4) Variational Bayesian framework - implemented in the program fastSTRUCTURE[56]. The program fastSTRUCTURE was run with a prior of 1 to 10 subgroups in the collection (K=1 to 10), and the output was parsed with "choosing model complexity" script to determine the possible range of subgroups. The potential subgroups identified were then inspected with known pedigree information and coefficient of coancestry values generated for each genotype to precisely identify the number of subgroups in the collection. The results were visualized using a plot made in R. Furthermore population structure analysis was also performed using GEMs. The normalized expression counts of 39, 609 GEMs was transformed to $\log_2$ scale, and 1000 GEMs with highest variances across the collection were utilized to generate a Eucledian distance matrix followed by hierarchical clustering using Ward's linkage method in R. A confusion table was built to compare the performance of the five different approaches to the known partial pedigree information, and decipher the population structure in the collection.

3.5.6 Linkage disequilibrium (LD), and LD decay in the collection. Linkage disequilibrium in the collection was investigated using the LD statistic "r^2." The r^2 values were generated for all marker pairs located within and between transcripts using the options "--r2, inter-chr and --ld-window-r2 0.0" in PLINK. We further mapped the Apios transcripts to the *Phaseolus vulgaris* genome assembly (version 1.0) using gmap (2014-05-15.v3) with

the following settings: "--cross-species, --format=coords, --npaths=0, --chimera-margin=0, intronlength=10034, --totallength=60969," and retrieved the likely location information for ~81% of the SNPs. These marker locations were then utilized to investigate the decay of LD along each of the "pseudo-Apios" chromosomes, as well as across the genome, with the assumption that the genome structure is conserved between the two phaseoloid legumes *A. americana* and *P. vulgaris*. The background LD estimated as 90th percentile of the r^2 value of marker-pairs on different chromosomes, and the commonly used criteria's (r^2=0.1 and 0.2) across the genome, were used as a threshold to determine the LD decay in the collection. Haplotype blocks were estimated using the options "--blocks, no-pheno-req and --blocks-max-kb 2000" in PLINK. The pericentromeric start and end coordinate of each of the chromosome was obtained from *P. vulagris* genome browser hosted at the Legume Information System (legumeinfo.org).

3.5.7 Association analysis using SNP markers. The phenotypic dataset used for association analysis is previously reported in Belamkar et al.[9], and contains REML-based LS means of 20 phenotypic traits that include 10 aboveground, and 10 belowground measurements recorded on 52 genotypes in Ames, IA during 2011 and 2012. The genotypic dataset contains 58,154 SNPs of high-quality that are filtered for minor allele frequency ≤ 0.1 % and maximum missing percentage ≥ 10%. Population structure was accounted for either by including only familial relatedness, or familial relatedness together with subpopulations in the linear mixed models (LMMs). Association analysis was first performed in the software program TASSEL[57] by incorporating (1) familial relatedness matrix (generated in PLINK) in the linear mixed model as random effect; and (2) familial relatedness (as random effect) and subpopulation membership coefficients generated using fastSTRUCTURE as covariates.

Association analysis was also conducted using "GCTA: a tool for Genome-wide Complex Trait Analysis[58]" with the difference being familial relatedness matrix generated in GCTA, and either six or 10 principal components used to account for presence of subpopulations. The performance of each of the model was tested using quantile-quantile (QQ) plots generated using "qqman" package in R. Marker-trait associations with a *P*-value less than 0.0001 were considered as significant associations. None of the associations were significant after adjusting for multiple testing using Bonferroni correction, which is probably due to the large number of SNP markers used in this study. Hence, the threshold *P*-value ($P < 0.0001$) utilized in this study was based on distribution of *P*-values in the present study, and threshold *P*-values utilized in previous studies with similar population size, or SNP markers[36, 59]. Significant ($P < 0.0001$) marker-trait associations were tabulated and represented in Manhattan plots generated using "qqman" package in R. The marker-traits associations were further assessed by comparing the biological function of the transcript containing the SNP marker (derived from its annotation) to the associated phenotypic trait.

3.5.8 Association analysis using gene expression markers. Linear regression analysis was performed in R with 39,609 GEMs as dependent variables, and each of the 20 phenotypic traits as independent variables. The adjusted-r^2 and significance (*P*) values were recorded for each of the GEM, per trait. The associations were: (1) filtered by applying Bonferroni correction of $P<0.05$, which resulted in a new threshold of $P<0.0000013$ (0.05/39609). (2) The associations that passed the Bonferroni test were examined for assumptions of linear regression. FPKM normalized expression values were represented along Y-axis, and REML-based LS means for the phenotypic trait were plotted along the X-axis. The significant associations that passed both Bonferroni test, and met the assumptions were tabulated, and

displayed using Manhattan plots. The significant associations were further verified by comparing the associated phenotypic trait with the annotations of transcripts.

3.6 Acknowledgments

The authors are thankful to Dr. Randy C. Shoemaker for providing laboratory facilities; to the late Berthal D. Reynolds for his contributions to the Apios breeding program during 1984-1995; to Dr. Dennis Wollard for providing seed material for a subset of the genotypes used in this study; to V. Gautam Bhattacharya for facilitating obtaining of germplasm as well as consistent encouragement; and to Jody Hayes, Rebecca Nolan, Alex Gascho, and Joshua McCombs for their invaluable support during phenotypic data collection and harvest.

Author contributions

VB and SBC conceived and designed the experiments. VB and SRK harvested leaf tissue, and collected phenotype data; VB isolated RNA and performed quality inspections; ADF performed sequencing and ran the Alpheus^{TM} pipeline for variant calling; NTW performed installation, troubleshooting and maintenance of the packages on the servers for analyses; VB performed the data analyses; VB and SBC wrote the manuscript. All authors revised and approved the manuscript.

Competing financial interests

The authors declare no competing financial interests.

3.7 References

1. Ortiz-Ceballos AI, Aguirre-Rivera JR, Salgado-Garcia S, Ortiz-Ceballos G. Maize–Velvet Bean Rotation in Summer and Winter Milpas: A Greener Technology. *Agron J* **107**, 330-336 (2015).

2. Vega-Galvez A, Miranda M, Vergara J, Uribe E, Puente L, Martinez EA. Nutrition facts and functional potential of quinoa (Chenopodium quinoa willd.), an ancient Andean grain: a review. *J Sci Food Agric* **90**, 2541-2547 (2010).

3. Peterson A, Jacobsen S-E, Bonifacio A, Murphy K. A Crossing Method for Quinoa. *Sustainability* **7**, 3230-3243 (2015).

4. Wu G, Morris CF, Murphy KM. Evaluation of texture differences among varieties of cooked quinoa. *J Food Sci* **79**, S2337-2345 (2014).

5. Cox TS, Van Tassel DL, Cox CM, DeHaan LR. Progress in breeding perennial grains. *Crop and Pasture Science* **61**, 513-521 (2010).

6. DeHaan LR. Progress in developing Kernza wheatgrass as a perennial grain. *Water, Food, Energy & Innovation for a Sustainable World*, (2013).

7. Blackmon WJ, Reynolds BD. The crop potential of *Apios americana*-preliminary evaluations. *Hortscience* **21**, 1334-1336 (1986).

8. Reynolds BD, Blackmon WJ, Wickremesinhe E, Wells MH, Constantin RJ. Domestication of *Apios americana*. In: *Advances in new crops* (eds Janick J, Simon JE). Timber Press (1990).

9. Belamkar V, Wenger A, Kalberer SR, Bhattacharya VG, Blackmon WJ, Cannon SB. Evaluation of Phenotypic Variation in a Collection of Apios americana: An Edible Tuberous Legume. *Crop Sci* **55**, 712-726 (2015).

10. Parker MA. Relationships of bradyrhizobia from the legumes Apios americana and Desmodium glutinosum. *Applied and Environmental Microbiology* **65**, 4914-4920 (1999).

11. Seabrook JAE, Dionne LA. Studies on the genus Apios. I. Chromosome number and distribution of Apios americana and A. priceana. *Canadian Journal of Botany* **54**, 2567-2572 (1976).

12. Bruneau A, Anderson GJ. Reproductive biology of diploid and triploid *Apios americana* (Leguminosae). *American journal of botany* **75**, 1876-1883 (1988).

13. Westerkamp C, Paul H. Apios americana, a fly-pollinated papilionaceous flower? *Pl Syst Evol* **187**, 135-144 (1993).

14. Joly S, Bruneau A, Galloway L. Evolution of triploidy in *Apios americana* (Leguminosae) revealed by genealogical analysis of the histone H3-D gene. *Evolution* **58**, 284-295 (2004).

15. O'Rourke JA*, et al.* An RNA-Seq transcriptome analysis of orthophosphate-deficient white lupin reveals novel insights into phosphorus acclimation in plants. *Plant Physiol* **161**, 705-724 (2013).

16. Kudapa H*, et al.* Comprehensive transcriptome assembly of Chickpea (Cicer arietinum L.) using sanger and next generation sequencing platforms: development and applications. *PloS one* **9**, e86039 (2014).

17. Wu N*, et al.* De novo next-generation sequencing, assembling and annotation of *Arachis hypogaea* L. Spanish botanical type whole plant transcriptome. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* **126**, 1145-1149 (2013).

18. Dubey A*, et al.* Defining the transcriptome assembly and its use for genome dynamics and transcriptome profiling studies in pigeonpea (*Cajanus cajan* L.). *DNA Res* **18**, 153-164 (2011).

19. Farrell JD, Byrne S, Paina C, Asp T. De novo assembly of the perennial ryegrass transcriptome using an RNA-Seq strategy. *PloS one* **9**, e103567 (2014).

20. Zou X*, et al.* Genome-wide single nucleotide polymorphism and Insertion-Deletion discovery through next-generation sequencing of reduced representation libraries in common bean. *Molecular Breeding* **33**, 769-778 (2014).

21. Hyten D*, et al.* High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC genomics* **11**, 38 (2010).

22. Hyten D*, et al.* High-throughput SNP discovery and assay development in common bean. *BMC genomics* **11**, 475 (2010).

23. Lai K, *et al.* Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant biotechnology journal* **10**, 743-749 (2012).

24. Van Tassell CP, *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Meth* **5**, 247-252 (2008).

25. Hyten DL, *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences* **103**, 16666-16671 (2006).

26. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature reviews Genetics* **11**, 800-805 (2010).

27. Cannon SB, *et al.* Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular biology and evolution* **32**, 193-210 (2015).

28. Doyle J. Polyploidy in Legumes. In: *Polyploidy and Genome Evolution* (eds Soltis PS, Soltis DE). Springer Berlin Heidelberg (2012).

29. Lucas MR, Diop N-N, Wanamaker S, Ehlers JD, Roberts PA, Close TJ. Cowpea–Soybean Synteny Clarified through an Improved Genetic Map. *Plant Gen* **4**, 218-225 (2011).

30. Kang YJ, *et al.* Draft genome sequence of adzuki bean, Vigna angularis. *Sci Rep* **5**, (2015).

31. Schmutz J, *et al.* A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics* **46**, 707-713 (2014).

32. Bisognin DA. Breeding vegetatively propagated horticultural crops. *Crop Breed Appl Biotechnol* **11**, 35-43 (2011).

33. Flint-Garcia SA, Thornsberry JM, Buckler ESt. Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* **54**, 357-374 (2003).

34. Abdurakhmonov IY, Abdukarimov A. Application of association mapping to understanding the genetic diversity of plant germplasm resources. *Int J Plant Genomics* **2008**, 574927 (2008).

35. Kadri NK, Guldbrandtsen B, Sorensen P, Sahana G. Comparison of genome-wide association methods in analyses of admixed populations with complex familial relationships. *PloS one* **9**, e88926 (2014).

36. Harper AL, *et al.* Associative transcriptomics of traits in the polyploid crop species Brassica napus. *Nature biotechnology* **30**, 798-802 (2012).

37. Fujii H, Verslues PE, Zhu JK. Identification of two protein kinases required for abscisic acid regulation of seed germination, root growth, and gene expression in Arabidopsis. *The Plant cell* **19**, 485-494 (2007).

38. Kulik A, Wawer I, Krzywinska E, Bucholc M, Dobrowolska G. SnRK2 protein kinases--key regulators of plant response to abiotic stresses. *OMICS* **15**, 859-872 (2011).

39. McConnell JR, Emery J, Eshed Y, Bao N, Bowman J, Barton MK. Role of *PHABULOSA* and *PHAVOLUTA* in determining radial patterning in shoots. *Nature* **411**, 709-713 (2001).

40. Aguilera-Carbó A, Montañez JC, Anzaldúa-Morales A, Reyes ML, Contreras-Esquivel J, Aguilar CN. Improvement of color and limpness of fried potatoes by in situ pectinesterase activation. *Eur Food Res Technol* **210**, 49-52 (1999).

41. Montañez Sáenz J, Téllez A, de la Garza H, de la Luz Reyes M, Contreras-Esquivel JC, Aguilar CN. Purification and some properties of pectinesterase from potato (Solanum tuberosum L.) alpha cultivar. *Brazilian Archives of Biology and Technology* **43**, 0-0 (2000).

42. Fernie AR, Willmitzer L. Molecular and Biochemical Triggers of Potato Tuber Development. *Plant Physiology* **127**, 1459-1465 (2001).

43. Sonnewald U, Hajirezaei M-R, Kossmann J, Heyer A, Trethewey RN, Willmitzer L. Increased potato tuber size resulting from apoplastic expression of a yeast invertase. *Nat Biotech* **15**, 794-797 (1997).

44. Licausi F*, et al.* HRE-type genes are regulated by growth-related changes in internal oxygen concentrations during the normal development of potato (*Solanum tuberosum*) tubers. *Plant & cell physiology* **52**, 1957-1972 (2011).

45. Ross EM, Moate PJ, Marett LC, Cocks BG, Hayes BJ. Metagenomic predictions: from microbiome to complex health and environmental phenotypes in humans and cattle. *PloS one* **8**, e73056 (2013).

46. Xu S, Zhang Q. Predicting yield of hybrid rice using omics data. In: *Genomic selection and genome-wide association studies* (ed^(eds Yu J, Garrick DJ). Plant and Animal Genome (2015).

47. Pruski K, Duplessis P, Lewis T, Astatkie T, Nowak J, Struik PC. Jasmonate effect on in vitro tuberization of potato (*Solanum tuberosum* L.) cultivars under light and dark conditions. *Potato Res* **44**, 315-325 (2001).

48. Nookaraju A*, et al.* Role of Ca2+-mediated signaling in potato tuberization: An overview. *Botanical Studies* **53**, 177-189 (2012).

49. Haas BJ*, et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protocols* **8**, 1494-1512 (2013).

50. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).

51. Miller NA*, et al.* Management of High-Throughput DNA Sequencing Projects: Alpheus. *Journal of computer science and systems biology* **1**, 132-132 (2008).

52. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635 (2007).

53. Stevens EL, Heckenberg G, Roberson ED, Baugher JD, Downey TJ, Pevsner J. Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS genetics* **7**, e1002287 (2011).

54. Lee T-H, Guo H, Wang X, Kim C, Paterson A. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC genomics* **15**, 162 (2014).

55. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82 (2011).

56. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Datasets. *Genetics*, (2014).

57. Zhang Z*, et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature genetics* **42**, 355-360 (2010).

58. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics* **46**, 100-106 (2014).

59. Hwang E-Y*, et al.* A genome-wide association study of seed protein and oil content in soybean. *BMC genomics* **15**, 1 (2014).

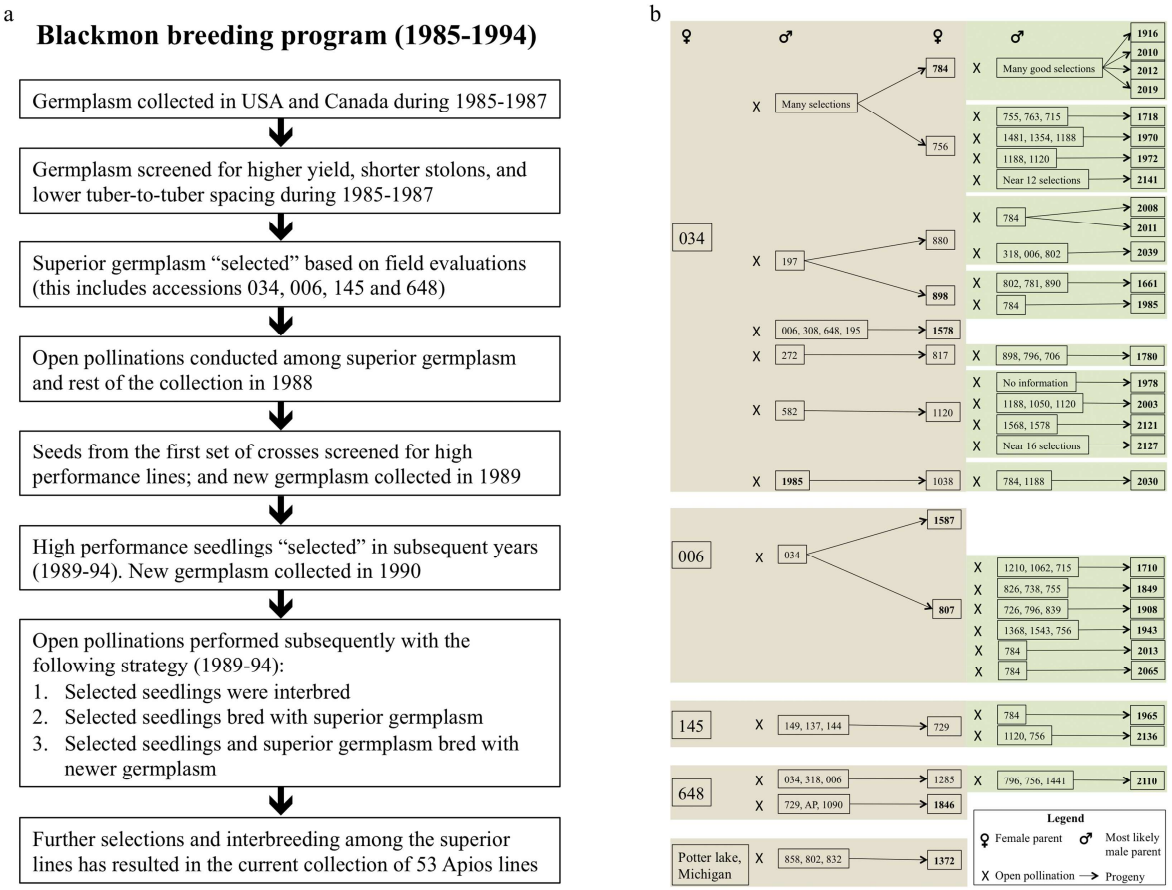

Figure 1 - Breeding strategy and pedigree of the *Apios americana* collection. (a) Breeding strategy utilized by Dr. Blackmon and Mr. Reynolds during 1985-1994 at Louisiana State University Agricultural Experiment Station in Baton Rouge, LA, for developing elite genotypes uses in this study. This information was interpreted from the field books of Dr. Blackmon and Mr. Reynolds. (b) Partial maternal lineage information that was traced from the field books for 35 of the 53 genotypes in the collection. The genotypes that produced seeds were recorded as maternal parents, and the genotypes that were growing close to the maternal parent were recorded as likely paternal parents. Hence, there are usually multiple paternal parents listed for a line. The genotypes in bold are the ones that exist in the collection used in this study.

Figure 2 - Heat map of the normalized RNA-Seq data showing expression of transcripts in six tissues of accession 2127. The normalized RNA-Seq data is in $\log_2$ scale. One thousand transcripts with highest variances across the 11 samples were utilized to make the heat map. Letters R1 and R2 represent replicates 1 and 2 of the corresponding tissues.

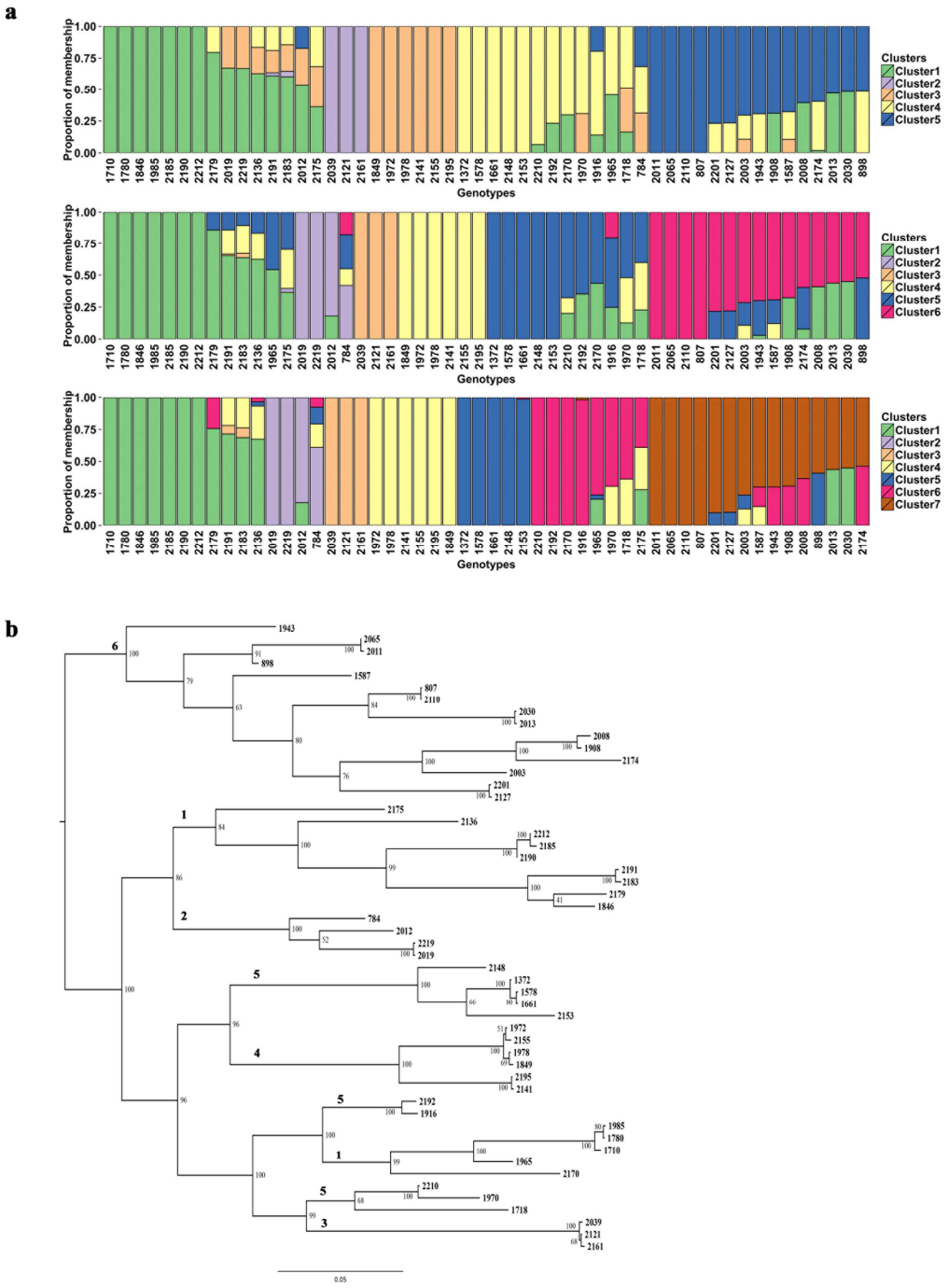

Figure 3 - Population structure in the collection. (a) Population structure using variational Bayesian framework - implemented in the program fastSTRUCTURE. The possible 5, 6 or 7 clusters (K=5, 6 or 7) identified in the collection are shown. The Y-axis represents proportion of membership of a genotype to the respective cluster, and X-axis indicates genotypes in the collection. (b) Phylogeny built using maximum likelihood implemented in the package SNPhylo. Letters 1 to 6 represent the clusters identified in Fig.3a.

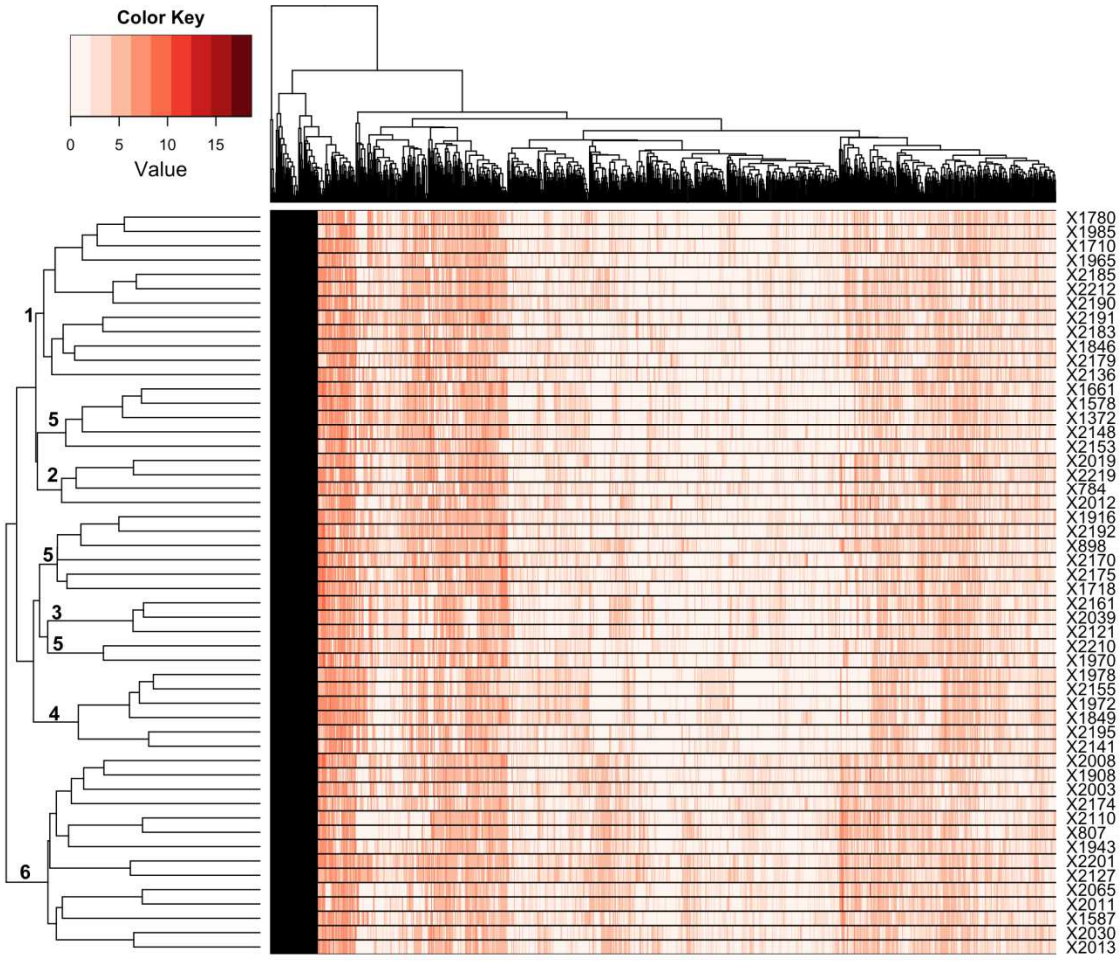

Figure 4 - Population structure in the collection using gene expression markers (GEMs). Phylogeny built using 1,000 GEMs (in $\log_2$ scale) that show highest variances across the 52 samples. A Euclidean distance matrix was utilized followed by hierarchical clustering with Ward's linkage. Letters 1 to 6 represent the clusters identified in Fig. 3a.

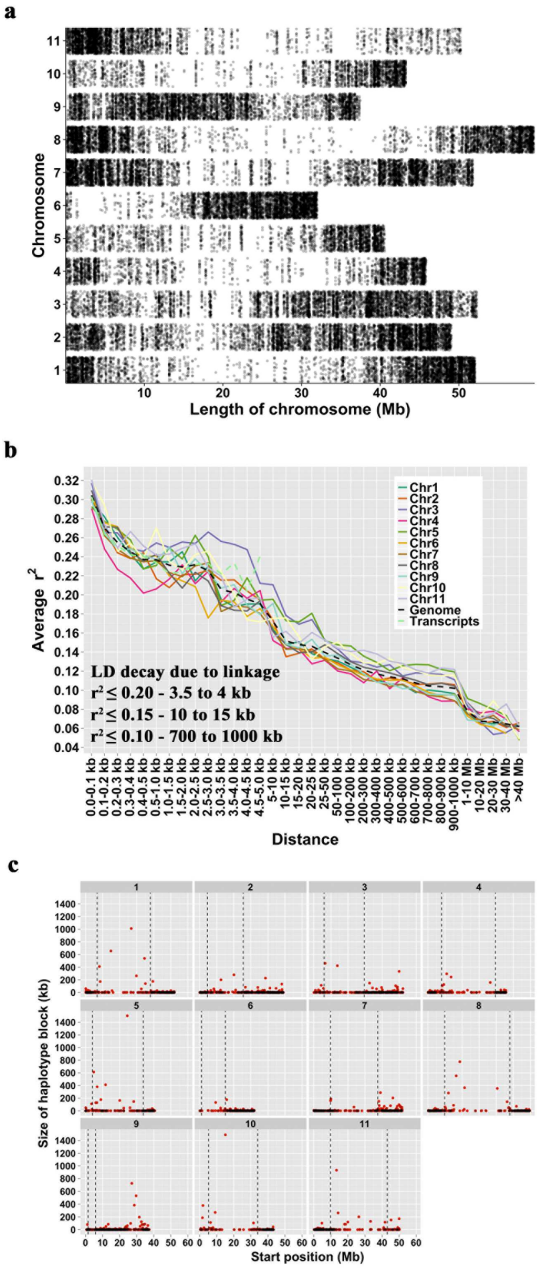

Figure 5 - Genome-wide SNP distribution, linkage disequilibrium and haplotype blocks. (a) Distribution of SNPs identified in the Apios collection along the 11 *Phaseolus vulgaris* chromosomes. Apios transcripts were mapped to the *Phaseolus vulgaris* genome assembly (version 1.0), and location information was retrieved for 46,852 of the 58,154 SNP makers. (b) Decay of linkage disequilibrium along each of the putative chromosomes, across the genome, and transcripts. (c) Distribution of haplotype blocks along each of the chromosome. The black dashed lines represent pericentromeric start and end coordinate of each of the chromosome obtained from *P. vulagris* genome.

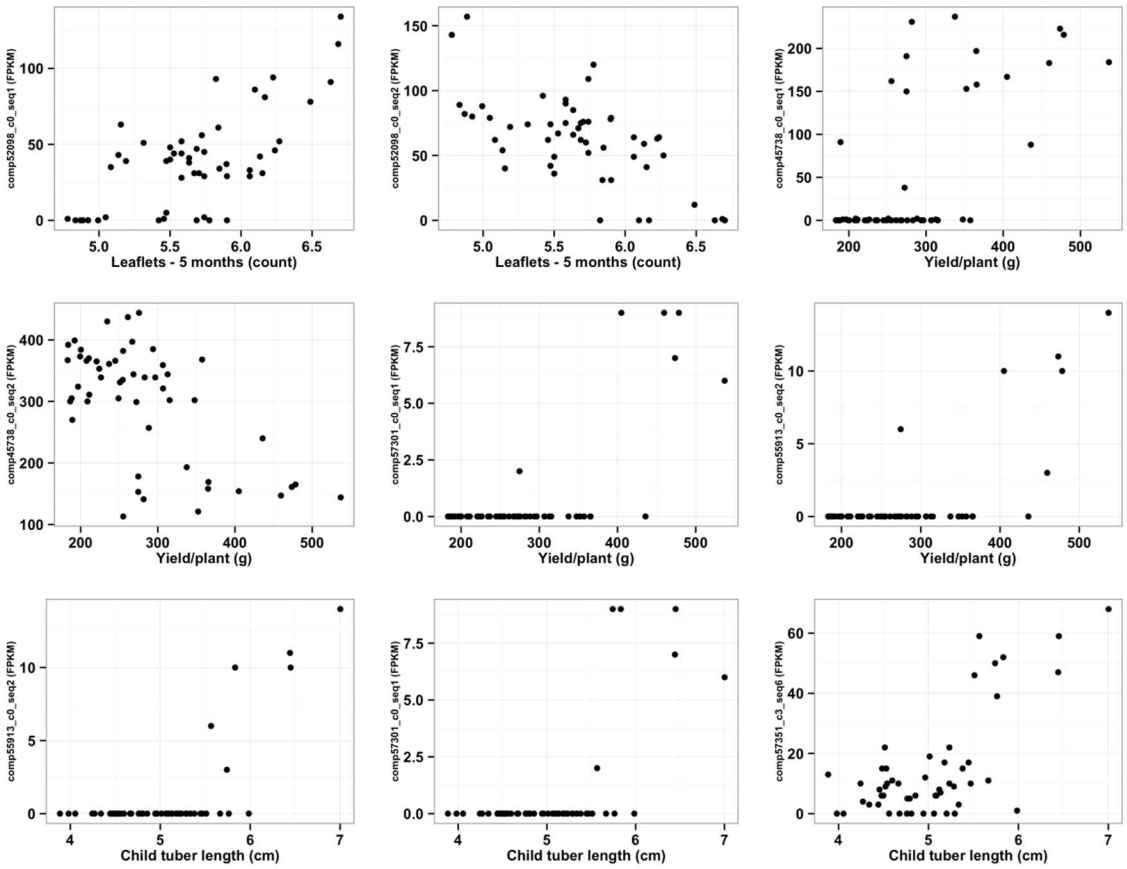

Figure 6 - Scatter plots of nine interesting marker-trait associations identified using gene expression markers (GEMs). Linear regression analysis was performed with 39,609 GEMs as dependent variables, and each of the 20 phenotypic traits as independent variables. We identified 28 GEM-trait associations, nine of which are shown in this figure. The Y-axis of each plot represents FPKM normalized expression values of a transcript for the 52 genotypes in the collection, and X-axis represents REML-based LS means for the phenotypic trait measured on the 52 genotypes.

Table 1. Statistics of the *Apios americana* transcriptome assembly built using accession 2127

Metric	Count
Sequencing reads	
Total number of reads	210,018,551
Total number of nucleotides (Gbp)	11.7
Total number of nucleotides used (Gbp)	11.5
Total number of nucleotides unused (Gbp)	0.2
***De novo* transcriptome assembly**	
Number of putative transcripts	96,560
Number of components (generally correspond with genes)	48,615
Number of components with splice variants	11,215
N50	1,863
Total number of nucleotides in the assembly (bp)	113,238,654
Peptides	
Number of peptides (alternative splice variants included)	60,880
Number of peptides (alternative splice variants excluded)	23,691
Annotation of peptides (alternative splice variants excluded)	
Swiss-Prot	16,711
Pfam-A	17,087
Signal peptide	2,240
Transmembrane helices	5,847
eggNOG (evolutionary genealogy of genes)	13,524
Gene Ontology	15,647
Expression	
Number of transcripts expressed in at least one tissue	56,735
Number of components expressed in at least one tissue	28,738

Table 2. Discovery of variants across 52 genotypes in the collection

Metric	Count
Number of accessions in the Apios collection	52
Total number of reads generated	1,315,730,442
Average number of reads generated per accession	25,302,509
Average % of reads mapped to the assembly per accession	89.5
Average % of reads mapped uniquely to the assembly per accession	61.4
Number of variants (SNP and Indels) identified with base criteria	1,582,730
Number of variants identified (base criteria + read depth ≥5)	299,145
Number of SNPs identified (base criteria + read depth ≥5)	271,170
Number of SNPs (base criteria + read depth ≥5 + MAF ≥0.1 + Maximum missing percentage ≤10%)	58,154
Average % reproducibility of SNPs (tested using 4 biological replicates)	90.6
Number of components (a.k.a genes) harboring SNPs	9,338
% of components in the assembly harboring SNPs	19.2
Average accessions missing per SNP (%)	1.5
Average heterozygous accessions per SNP (%)	37.5
Number of gene expression markers identified	39,609

Base criteria: At least 2 reads calling variant within at least 1 accession; ≥ 20% of the reads calling the variant allele in that sample; and average quality of bases calling the variant is ≥ 10)

Table 3. Marker-trait associations identified in the collection using SNP-based association analysis

Trait	SNP	Reference transcript	Position	Chr[1]	Position[2]	Alleles[3]	Freq[4]	Effect[5]	SE[6]	*P*-value	Annotation
First leaf emergence	S_23737486	comp54771_c0_seq1	1,673	04	43,022,415	A/G	0.23	0.16	0.04	8.71E-05	Serine/threonine protein kinase
Ground to first leaf	S_4036060	comp43747_c0_seq1	1,279	NA[7]	NA	C/T	0.21	1.35	0.33	3.84E-05	Succinyl-CoA ligase [ADP-forming] subunit alpha-2
Ground to first leaf	S_8354256	comp48670_c0_seq1	510	09	21,384,837	T/C	0.18	1.65	0.39	2.12E-05	Ubiquitin-fold modifier-conjugating enzyme
Ground to first leaf	S_12390026	comp50830_c0_seq2	355	01	269,531	T/C	0.16	1.39	0.35	5.47E-05	Glutathione S-transferase
Ground to first leaf	S_17227020	comp52764_c0_seq3	354	NA	NA	C/T	0.12	1.32	0.34	9.52E-05	PAP2 superfamily
Leaflets-2 months	S_11450514	comp50326_c0_seq1	552	10	43,192,696	A/T	0.50	-0.44	0.11	7.39E-05	StAR-related lipid transfer protein
Plant vigor	S_23474932	comp54688_c0_seq2	1,435	03	30,967,083	G/A	0.15	0.39	0.09	3.26E-05	NA
Stem diameter-5 months	S_30601377	comp56353_c0_seq1	1,880	02	47,724,772	A/T	0.33	-0.35	0.09	9.09E-05	Synaptotagmin-5
SPAD	S_31213411	comp56475_c0_seq1	2,559	05	14,781,013	G/T	0.46	0.97	0.25	9.82E-05	Nuclear pore complex protein Nup98-Nup96
Yield/plant	S_806532	comp29700_c0_seq1	260	NA	NA	T/C	0.20	97.65	23.40	3.02E-05	Glutaredoxin-C2
Tubers/plant	S_25048846	comp55124_c0_seq4	465	05	9,906,844	T/C	0.46	-9.91	2.54	9.73E-05	NA
Tubers/plant	S_41908941	comp58209_c0_seq1	117	01	50,369,688	T/A	0.48	-14.07	3.23	1.32E-05	Clathrin heavy chain 2; endocytosis
Stolon length	S_7515611	comp48154_c1_seq1	805	05	43,471	T/C	0.27	-9.89	2.54	9.78E-05	Deoxynucleoside kinases
Stolon length	S_14498332	comp51702_c0_seq1	1,505	03	40,709,216	G/A	0.24	-10.22	2.55	6.35E-05	RNA processing and splicing; WW domain-binding protein
Stolon length	S_17413318	comp52828_c0_seq1	183	NA	NA	T/A	0.26	-9.05	2.19	3.71E-05	NA
Mother tuber weight	S_18970167	comp53339_c0_seq2	904	08	57,655,132	T/C	0.48	-85.81	20.86	3.88E-05	Glucose-6-phosphate transmembrane transporter
Mother tuber length	S_25933856	comp55332_c0_seq1	1,714	02	499,012	G/A	0.38	0.79	0.20	6.22E-05	Regulator of nonsense transcripts
Mother tuber width	S_23095079	comp54598_c3_seq1	295	10	39,313,260	C/T	0.19	-0.90	0.23	9.28E-05	Pectinesterase
Child tuber weight	**S_12121278**	comp50661_c0_seq1	1,089	02	45,978,408	A/C	0.12	15.30	3.59	2.00E-05	PC-Esterase GDSL/SGNH-like Acyl-Esterase
Child tuber weight	S_28587045	comp55916_c1_seq1	1,482	08	7,701,745	T/A	0.15	11.83	3.03	9.55E-05	BTB/POZ domain-containing protein; signal transduction
Child tuber length	**S_12121278**	comp50661_c0_seq1	1,089	02	45,978,408	A/C	0.12	0.96	0.24	6.62E-05	PC-Esterase GDSL/SGNH-like Acyl-Esterase
Child tuber length	S_30406303	comp56304_c4_seq2	1,682	06	29,252,859	G/A	0.13	0.74	0.19	7.84E-05	Zinc finger CCCH; DNA binding; photomorphogenesis

[1]Chr represents chromosomal location of SNP on *Phaseolus vulgaris* genome; [2]Position represents location of SNP on *Phaseolus vulgaris* genome; [3]Effective allele/other allele at the SNP location; [4]Frequency of effective allele in the collection; [5]SNP effect corresponding to the effective allele; [6]Standard error associated with the SNP effect; [7]NA represents information not available. SNPs associated with multiple traits are in bold.

Table 4. Marker-trait associations identified in the collection using gene expression markers

Trait	Transcript	Chr[1]	Start[2]	Estimate[3]	SE[4]	t-value[5]	*P*-value	Adj-r[2]	Annotations
Internode length	comp56323_c0_seq5	NA[6]	NA	-2.74	0.46	-5.99	2.26E-07	0.41	NA
Stem diameter - 2 months	**comp57243_c2_seq2**	02	3,455,626	73.48	13.0	5.65	7.57E-07	0.38	RNA-binding protein 8A; mRNA processing
	comp52044_c2_seq3	10	42,053,143	-13.38	2.35	-5.69	6.68E-07	0.38	Heat shock 70 kDa protein 15/Molecular chaperone
Leaflets - 2 months	comp49135_c0_seq1	10	6,001,220	11.10	1.56	7.10	4.24E-09	0.49	CER1-like 1/Fatty acid hydroxylase
	comp55774_c3_seq6	05	34,676,406	-4.11	0.72	-5.72	5.92E-07	0.38	NA
Leaflets - 5 months	comp52098_c0_seq1	02	33,390,379	45.40	6.89	6.59	2.59E-08	0.45	Geranylgeranyl pyrophosphate synthase
	comp52098_c0_seq2	02	33,390,379	-45.56	7.62	-5.98	2.35E-07	0.41	Geranylgeranyl pyrophosphate synthase
Yield/plant	**comp57301_c0_seq1**	08	51,938,872	0.02	0.00	6.96	6.96E-09	0.48	GDSL esterase/lipase At1g71250; Geranylgeranyl reductase, choloroplastic
	comp55913_c0_seq2	08	52,258,272	0.03	0.00	6.70	1.80E-08	0.46	Integral membrane protein; Late exocytosis, Golgi transport
	comp49849_c0_seq1	05	30,357,198	0.04	0.01	6.48	3.93E-08	0.45	NA; Possesses a signal peptide cleavage site
	comp45738_c0_seq1	05	28,501,950	0.64	0.11	6.11	1.50E-07	0.42	14-3-3-like protein B
	comp32096_c0_seq1	05	20,713,186	-0.03	0.01	-5.76	5.17E-07	0.39	Uncharacterized protein At1g18480; Calcineurin-like phosphoesterase
	comp55939_c2_seq5	03	23,432,180	0.01	0.00	5.73	5.83E-07	0.38	UDP-glucose flavonoid 3-O-glucosyltransferase 7
	comp45738_c0_seq2	05	28,501,950	-0.69	0.12	-5.57	1.01E-06	0.37	14-3-3-like protein B
Tubers/plant	comp52843_c0_seq8	07	38,262,470	0.46	0.07	6.25	8.83E-08	0.43	ADP-ribosylation factor GTPase-activating protein AGD7

Table 4 continued

	comp57897_c0_seq2	08	4,984,724	0.60	0.11	5.66	7.42E-07	0.38	N-terminal kinase-like protein
	comp53873_c0_seq2	08	9,973,071	0.40	0.07	5.58	9.71E-07	0.37	Ferrochelatase-2, chloroplastic
	comp46887_c0_seq3	09	21,886,278	0.24	0.04	5.58	9.94E-07	0.37	NA
	comp54966_c0_seq3	11	8,691,266	-0.94	0.17	-5.56	1.06E-06	0.37	Nucleoside-diphosphate-sugar epimerases; UDP-glucuronic acid oxidase
Mother tuber length	comp57964_c0_seq4	07	41,689,316	-4.11	0.68	-6.07	1.70E-07	0.41	Mannosylglycoprotein endo-beta-mannosidase; Carbohydrate metabolism
Child tuber weight	**comp55913_c0_seq2**	08	52,258,272	0.21	0.04	5.75	5.34E-07	0.39	Integral membrane protein; Late exocytosis, Golgi transport
Child tuber length	**comp55913_c0_seq2**	08	52,258,272	3.41	0.50	6.79	1.26E-08	0.47	Integral membrane protein; Late exocytosis, Golgi transport
	comp57351_c3_seq6	02	47,947,761	19.12	2.83	6.76	1.42E-08	0.47	Jasmonate ZIM domain-containing protein 3; JA signaling pathway
	comp55939_c2_seq5	03	23,432,180	1.72	0.29	5.93	2.80E-07	0.40	UDP-glucose flavonoid 3-O-glucosyltransferase 7
	comp56129_c4_seq6	NA	NA	8.05	1.36	5.92	2.87E-07	0.40	NA
	comp56118_c0_seq1	02	43,809,153	2.12	0.37	5.72	6.02E-07	0.38	NA
	comp57301_c0_seq1	08	51,938,872	2.38	0.42	5.71	6.13E-07	0.38	GDSL esterase/lipase At1g71250; Geranylgeranyl reductase, choloroplastic
	comp57243_c2_seq2	02	3,455,626	31.05	5.45	5.70	6.40E-07	0.38	RNA-binding protein 8A; mRNA processing

[1]Chr represents chromosomal location of transcript on *Phaseolus vulgaris* genome; [2]Start represents start position of the transcript mapped to *Phaseolus vulgaris* genome; [3]Estimate obtained from linear regression analysis; [4]Standard error associated with the estimate; [5]t-value is from a test with null hypothesis that the estimate is equal to zero (no effect); [6]NA represents information not available. Transcripts associated with multiple traits are in bold.

Additional Files

Figure S1. Transcript length distribution of the 96,560 transcripts in the *Apios americana de novo* assembly.

Figure S2. Sequence conservation of transcripts inspected by performing a BLASTX search with a threshold of 1E-05 against the proteomes of six related legumes and *Arabidopsis thaliana*. Gray: percentage of transcripts in the assembly matching peptides; black: percentages of peptides in the respective species matching Apios transcripts.

Figure S3. Negative correlation between inbreeding coefficients estimated for each genotype and tubers produced per plant recorded in Ames, IA during 2011-2012. A linear regression was performed with tubers/plant as response variable and inbreeding coefficients as independent variables. The estimate was -38.57, standard error 10.13, t-value -3.807, P=0.000386, and the adjusted R-squared was 0.21.

Figure S4. Phylogeny built using a distance matrix (1-IBS) based on Identity-by-state (IBS), and performing a hierarchical clustering using Ward's linkage. Letters 1 to 6 represent the clusters identified in Fig. 3a.

Figure S5. Principal component analysis. PC1 and PC2 represent principal component 1 and 2 respectively.

Figure S6. Number of Apios SNPs on each of the 11 *P. vulgaris* chromosomes. Nearly 81% of the SNPs indentified in the Apios collection were mapped to common bean genome. The mean and median numbers of SNPs per chromosome are 4,259 and 4,592, respectively.

Figure S7. Quantile-Quantile (QQ) plots for evaluating performances of the different models and software packages utilized for association anlaysis. Association analysis was first performed in the software program TASSEL by incorporating (1) familial relatedness matrix (kinship; generated in PLINK) in the linear mixed model as random effect; and (2) familial relatedness (as random effect) and subpopulation membership coefficients generated using fastSTRUCTURE as covariates. Association analysis was also conducted using "GCTA: a tool for Genome-wide Complex Trait Analysis" with the difference being familial relatedness matrix generated in GCTA, and either six or 10 principal components used to account for presence of subpopulations.

Figure S8. Manhattan plots displaying marker-trait associations identified in the collection. (a) Marker-trait associations with the aboveground traits. (b) Marker-trait associations with the belowground traits.

Figure S9. Gene expression markers (GEMs) associated with aboveground measurements and yield/plant after correcting for multiple-testing using Bonferroni correction of P<0.05 (new threshold P<0.0000013). Each scatter plot corresponds to a GEM associated with a trait. The Y-axis represents normalized expression value of GEM in the 52 genotypes, and X-axis corresponds to the phenotypic measurement of the 52 genotypes for the respective trait.

Figure S10. Gene expression markers (GEMs) associated with belowground measurements (except yield/plant) after correcting for multiple-testing using Bonferroni correction of *P*<0.05 (new threshold *P*<0.0000013). Each scatter plot corresponds to a GEM associated with a trait. The Y-axis represents normalized expression value of GEM in the 52 genotypes, and X-axis corresponds to the phenotypic measurement of the 52 genotypes for the respective trait.

Figure S11. Manhattan plots displaying gene expression markers associated with the aboveground traits.

Figure S12. Manhattan plots displaying gene expression markers associated with the belowground traits.

Figure S13. Linkage disequilibrium decay between the two transcripts "comp57301_c0_seq1 and comp55913_c0_seq2." The dashed line at $r^2 = 0.15$ corresponds to background LD in the genome.

Table S1. Genome size estimates of 25 genotypes from Blackmon-Reynolds collection and wild samples estimated using flow cytometry

Table S2. Summary of RNA-Seq data and expression in different tissues of genotype 2127

Table S3. Validation of single nucleotide polymorphisms (SNPs) using biological replicates

Table S4. Frequency of alleles in the SNP dataset used in this study

Table S5. Correlations between inbreeding coefficients estimated for each genotype and phenotypic measurements recorded in Ames, IA in 2011-2012

Table S6. Parent-child, and half-sib relationships (0.46>=PI_HAT<=0.54) identified in the Apios collection

Table S7. Comparison of five different approaches of population structure analyses with known maternal pedigree information available from Blackmon-Reynolds breeding program

Table S8. Linkage disequilibrium decay along the chromosomes, whole genome and transcripts at different r2 thresholds

Table S9. Potential donor accessions based on SNP based marker-trait associations identified in the collection

CHAPTER 4. COMPREHENSIVE CHARACTERIZATION AND RNA-SEQ PROFILING OF THE HD-ZIP TRANSCRIPTION FACTOR FAMILY IN SOYBEAN (*GLYCINE MAX*) DURING DEHYDRATION AND SALT STRESS

A paper published in BMC Genomics 2014, 15:950

The electronic version of this article can be found online at:

http://www.biomedcentral.com/1471-2164/15/950

Vikas Belamkar[1,2], Nathan T Weeks[3], Arvind K Bharti[4], Andrew D Farmer[4], Michelle A Graham[2,3] and Steven B Cannon[2,3]

[1] Interdepartmental Genetics, Iowa State University, Ames, IA 50011, USA

[2] Department of Agronomy, Iowa State University, Ames, IA 50011, USA

[3] United States Department of Agriculture - Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA

[4] National Center for Genome Resources, Santa Fe, NM 87505, USA

4.1 Abstract

Background

The homeodomain leucine zipper (HD-Zip) transcription factor family is one of the largest plant specific superfamilies, and includes genes with roles in modulation of plant growth and response to environmental stresses. Many HD-Zip genes are characterized in

Arabidopsis (*Arabidopsis thaliana*), and members of the family are being investigated for abiotic stress responses in rice (*Oryza sativa*), maize (*Zea mays*), poplar (*Populus trichocarpa*) and cucumber (*Cucmis sativus*). Findings in these species suggest HD-Zip genes as high priority candidates for crop improvement.

Results

In this study we have identified members of the HD-Zip gene family in soybean cv. 'Williams 82', and characterized their expression under dehydration and salt stress. Homology searches with BLASTP and Hidden Markov Model guided sequence alignments identified 101 HD-Zip genes in the soybean genome. Phylogeny reconstruction coupled with domain and gene structure analyses using soybean, *Arabidopsis*, rice, grape (*Vitis vinifera*), and *Medicago truncatula* homologues enabled placement of these sequences into four previously described subfamilies. Of the 101 HD-Zip genes identified in soybean, 88 exist as whole-genome duplication-derived gene pairs, indicating high retention of these genes following polyploidy in *Glycine* ~10 Mya. The HD-Zip genes exhibit ubiquitous expression patterns across 24 conditions that include 17 tissues of soybean. An RNA-Seq experiment performed to study differential gene expression at 0, 1, 6 and 12 hr soybean roots under dehydration and salt stress identified 20 differentially expressed (DE) genes. Several of these DE genes are orthologs of genes previously reported to play a role under abiotic stress, implying conservation of HD-Zip gene functions across species. Screening of HD-Zip promoters identified transcription factor binding sites that are overrepresented in the DE genes under both dehydration and salt stress, providing further support for the role of HD-Zip genes in abiotic stress responses.

Conclusions

We provide a thorough description of soybean HD-Zip genes, and identify potential candidates with probable roles in dehydration and salt stress. Expression profiles generated for all soybean genes, under dehydration and salt stress, at four time points, will serve as an important resource for the soybean research community, and will aid in understanding plant responses to abiotic stress.

Keywords: Soybean; HD-Zip; Transcription factor; Gene family; Whole-genome duplication; RNA-Seq; Dehydration stress; Salt stress; Abiotic stress.

4.2 Background

Plants sense and respond to environmental variations in temperature, nutrient availability, water level, and light conditions. The homeodomain leucine zipper (HD-Zip) transcription factors play a significant role in regulating plant growth adaptation responses by integrating developmental and environmental signals. Homeodomain leucine zipper (HD-Zip) transcription factors have been found exclusively in the plant kingdom [1,2], the only exception being the recent identification in the charophycean algae [3]. The characteristic feature of the HD-Zip gene family is the association of homeodomain (HD) and the leucine zipper (LZ) motif in a single protein. In other kingdoms, they are present as domains of distinct proteins. The homeodomain is a ~60 amino acid DNA binding domain composed of a helix-turn-helix structure that folds into three characteristic alpha-helices, capable of interacting specifically with DNA [2]. The LZ motif is a dimerization motif and is located

immediately after the HD. The LZ motif allows the formation of homo- and hetero-dimers that are required for binding to DNA. The HD-Zip transcription factors can be subdivided into four subfamilies: HD-Zip I to IV, based on distinct sequence features (DNA-binding domains and additional conserved motifs that are specific to each of the subfamilies), and distinct functions of proteins from each of the subfamilies (for reviews, see [1,4]).

The HD-Zip superfamily has been analyzed in several species including *Arabidopsis* (*Arabidopsis thaliana*) [1,5,6], rice (*Oryza sativa*) [4,7], maize (*Zea mays*) [8], poplar (*Populus trichocarpa*) [9], and the HD-Zip I and IV genes in cucumber (*Cucumis sativus*) [10,11]. However, functional characterization studies have been limited to the model plant *Arabidopsis*, while a few selected genes have been investigated in other species [1,4]. A subset of the HD-Zip genes have recently been described in soybean (*Glycine max*) [12]. HD-Zip genes are involved in several abiotic stress responses, meristem regulation, photomorphogenesis, and root development [1,4]. The HD-Zip I genes have been investigated for their roles in water deficit and salt stress responses. The HD-Zip I *Arabidopsis* genes *ATHB7* and *ATHB12*, and their orthologs in other species, including *HaHB4* from sunflower (*Helianthus annuus*), *NaHD20* in *Nicotiana attenuatta*, and *OsHOX6* in rice (*Oryza sativa*), have increased expression under water-stress conditions [7,13-15]. ATHB7 and ATHB12 act as negative regulators of growth and development by reducing plant growth under water-deficit conditions [13,16,17]. *HaHB4* delays the onset of senescence when expressed in *Arabidopsis* [18,19]. The *Arabidopsis* HD-Zip I genes *ATHB5* and *ATHB6*, and the homologs in *Craterostigma plantagineum CpHB5*, *6*, *7* and *CpHB8*, are involved in water deficit stress [20]. ATHB5 acts as a positive regulator of ABA responsiveness at the seedling stage, with elevated levels of *ATHB5* resulting in higher ABA

responsiveness. On the contrary, ABA reduces the wild-type expression of *ATHB5*, indicating *ATHB5* is part of a negative feedback loop regulating ABA sensitivity in the germinating seedlings [21]. This implies ATHB5 mediates an initial response of the seedling to an ABA signal imposed (for instance, seedling development under limited-term water-deficit conditions) - but reduces the response to extended water stress. The *Arabidopsis* gene *ATHB6* has increased expression under water deficit stress [22]. *Arabidopsis* plants overexpressing ATHB6 display lowered stomatal closure and reduced inhibition of germination by ABA [23] - the characteristics of the ABA-insensitive mutant *abi1* and *abi2* [24]. Deng et al. [25] suggested that ATHB6 acts as a negative regulator of ABA response under water deficit stress.

A recent study in maize found all 17 HD-Zip I genes differentially expressed (DE) under drought stress [8]. The *Arabidopsis* genes *ATHB21*, *ATHB40* and *ATHB53* and the *M. truncatula* gene *MTHB1* are induced by salt stress [4]. The over-expression of MTHB1 reduces lateral root emergence. Ariel et al. [26] proposed reduced lateral root growth as a mechanism to reduce the exposure of the roots to high saline soil. The *Arabidopsis* gene *HAT22*, the rice genes *OsHOX11* and *OsHOX27*, and the *C. plantagineum* genes *CpHB1* and *CpHB2*, all in HD-Zip II, are induced by water stress [7,20,27]. Thus, it is evident that members of the HD-Zip I and II are enriched for genes that are involved in water deficit and salt stress. The emphasis in the literature has focused on HD-Zip I proteins for their role in abiotic stress, while systematic characterization of genes from the other subfamilies has been lacking. A recent study in rice shows the importance of investigating genes from other subfamilies. Yu et al. [28] demonstrated the overexpression of a HD-Zip IV gene (HDG11) confers drought tolerance, and increases yield under both normal and drought conditions.

With the utilization of high throughput sequencing techniques such as RNA-Seq, it is possible to investigate the expression of HD-Zip genes belonging to all subfamilies in the same experiment, and identify potential candidates for functional characterization studies.

The identification and classification of HD-Zip genes in prior studies has been based on homology searches, well-conserved domains and motifs in each of the subfamilies, and on conserved gene structures among subfamily members [5-11]. The availability of whole genome sequence information for increasing numbers of angiosperm species has enabled utilization of evolutionary relationships among the species to help characterize HD-Zip genes. Evolutionary relationships among species in a gene family analysis can be combined with whole genome duplication (WGD) histories. The eudicots *Arabidopsis*, grape, soybean and *M. truncatula* share a common "gamma" genome triplication event that occurred around 117 million years ago (Mya), early in the eudicot evolution [29,30]. The *Arabidopsis* lineage shows a signal for two additional rounds of WGD events within the last 70 million years [30,31]. Soybean and *Medicago* share a common legume-specific WGD event approximately 59 Mya [32,33], and soybean has undergone an additional glycine-specific genome duplication event around 13 Mya [30,32]. Rice shows evidence of two rounds of WGD events [30]. The grape genome has undergone a genome triplication event, but lacks a recent WGD event [34]. Conceptually, a single-copy gene in the ancestor of angiosperm plants and retained after every WGD event would give rise to the following numbers of homologous genes: 3 in grape, 6 in *Medicago,* 4 in rice, and 12 each in soybean and *Arabidopsis*. There is also evidence for two even older WGDs: one at around 320 Mya, prior to the separation of angiosperms and gymnosperms and referred to as the "ancestral seed plant WGD;" and another at around 190 Mya, predating the origin of angiosperms and termed the "ancestral

angiosperm WGD" [31]. Per this model of WGDs, an angiosperm gene family is typically comprised of four old angiosperm clades, assuming a starting point of one gene copy in the ancestor of seed plants. We examine the HD-Zip family in the context of this hypothesized history of WGDs, and provide insights into evolutionary history of each of the subfamilies relative to these WGD events.

In this study, we have 1) identified all putative HD-Zip genes in the soybean genome and placed them into their respective subfamilies; 2) provided phylogenetic relationships among HD-Zip proteins from eight species that include six eudicots: poplar, cucumber, *Arabidopsis*, grape, soybean and *M. truncatula*, and two moncots: rice and maize; 3) characterized the structures of all HD-Zip genes; 4) described the genomic organization, tracing the expansion of the gene family through WGD events; 5) presented gene expression data for all HD-Zip genes in 24 conditions including at least 17 different tissues of soybean; 6) provided RNA-Seq based gene expression profiles of all soybean genes including HD-Zip genes, in the roots under normal conditions and dehydration and salt stress after 0, 1, 6 and 12 hr treatments; and 7) identified genes that may participate in HD-Zip gene pathways by screening HD-Zip promoters for conserved motif of transcription factor binding sites (TFBSs).

4.3 Methods

4.3.1 Homology searches, multiple sequence alignments, and phylogenetic analysis

The sequences of 47 HD-Zip (17 HD-Zip I, 9 HD-Zip II, 5 HD-Zip III and 16 HD-Zip IV) proteins of *A. thaliana* (TAIR8_genome_release, 11/30/09) described in Ariel et al.

[1], were obtained from TAIR [35]. The proteomes of four other sequenced angiosperms *G. max* (assembly v1.01, JGI Glyma1.0 annotation), *M. truncatula* (v 3.5.1), *O. sativa* (MSU Release 6.0) and *V. vinifera* (12X March 2010 release) were obtained from the respective repositories for these genomes and BLAST databases were built for each of them on our local server. A BLASTP v2.2.22 (protein-protein BLAST) [36] search with a threshold of 1E-10 was used for initial identification of the homologous *Arabidopsis* HD-Zip genes in each of the genomes described above. The multiple sequence alignment of the homologous sequences from the five species was performed using MUSCLE v3.8.31 [37]. The alignment was manually inspected and trimmed using SeaView v4.2.5 [38,39] and BBEdit v8.7.2 respectively. A preliminary phylogenetic tree (not shown) encompassing four HD-Zip subfamilies was built using CLUSTAL v2.0.12 [40] and the tree was visually examined using FigTree v1.3.1 [41].

The probable HD-Zip genes belonging to each of the four subfamilies were identified based on the clustering of sequences with known HD-Zip genes from *Arabidopsis* in the preliminary phylogenetic tree. The outlier sequences that did not cluster with any *Arabidopsis* genes were temporarily excluded. The probable HD-Zip genes were then aligned using MUSCLE to build a profile Hidden Markov Model (HMM) separately for each subfamily using the hmmbuild program, implemented in the package HMMER v3.0b2 [42]. The probable HD-Zip sequences were re-aligned to the profile HMM using hmmalign, available in the tool HMMER, and were viewed in SeaView. Sequence logos were generated for each subfamily using the web tool WebLogo [43] to identify conserved regions in the alignments (Additional file 1: Figure S1, Additional file 2: Figure S2, Additional file 3: Figure S3 and Additional file 4: Figure S4). The HMM alignments were trimmed to retain

the conserved regions (HMM "match states") using BBEdit. The trimmed alignments were used to build the phylogenetic trees for each subfamily using the maximum likelihood method implemented in PhyML v3.0 [44] available at iPlant Collaborative [45] using default settings. The approximate likelihood ratio test (aLRT) branch support values [46] are displayed on the branches in percentages. The phylogenetic tree for each subfamily was displayed using FigTree. The rooting was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies (data not shown).

The outlier sequences excluded based on the preliminary phylogenetic tree were used in a search against HMM of each subfamily using the hmmpfam available in the tool HMMER v2.3.2 and the membership of sequences in each subfamily was investigated. The process of generating a phylogenetic tree followed by excluding outlier sequences, re-alignments, building HMMs, re-aligning using HMM, and rebuilding the trees, was iterated several times for each subfamily. A phylogenetic tree with appropriate tree topology based on evolutionary relationship among the five species was then generated for each subfamily. Lastly, we added HD-Zip I to IV sequences from maize and poplar, and HD-Zip I and IV sequences from cucumber that have recently been reported [8-11] to the final phylogenetic trees. This will allow the investigation of orthologous sequences from eight species that includes HD-Zip genes identified in all angiosperm species to date.

4.3.2 Validation, structural characterization, and duplication history of HD-Zip genes

The HD-Zip subfamilies have remarkably well-conserved domains, motifs, and gene structures [1,2,4] that can be utilized to validate genes identified using phylogenetic analysis. All sequences identified as HD-Zip genes as well as outlier sequences (excluded after preliminary phylogenetic tree construction) were used as queries in a batch search [47]

against Pfam 27.0, with an E-value threshold of 1E-3 to identify the conserved domains. The conserved motifs were investigated by examining the sequence logos that were generated using HMM sequence alignments of each subfamily. The gene structure was studied using the exon-intron organization in the pre-mRNA. The gene structures were rendered using the *G. max* cv. Williams 82 gene models (assembly v1.01, JGI Glyma1.0 annotation) that were downloaded from Phytozome [48] and using a modified version of the Bio-Graphics 2.25 feature_draw.pl script [49]. The genomic locations were obtained from the GFF file of *G. max* assembly v1.01, JGI Glyma1.0 annotation, and were displayed using chromosome visualization tool (CViT) [50]. The homoeologous HD-Zip gene pairs that are a result of the early-legume WGD event (~59 Mya), and the *Glycine*-specific duplication event (~13 Mya), were inferred from the phylogeny, as well as from the syntenic paralog pair information available for all soybean genes from the Joint Genome Institute (JGI) at Phytozome [51]. Paralogous genes resulting from tandem duplication events were identified based on their proximity on the same chromosome [52] and pairing in the same clade in the phylogenetic tree.

4.3.3 Expression profiles of HD-Zip genes in 24 conditions (17 tissues) of soybean

An RNA-Seq atlas of *G. max* describing expression of genes in 24 conditions including at least 17 different tissues of soybean was reported by Severin et al. [53] and Libault et al. [54,55]. The Reads/Kb/Million (RPKM) normalized data for 14 tissues investigated by Severin et al. are available for download and interactive analysis at SoyBase [56], and expression data for three additional tissues, and tissues infected with the bacterium *Bradyrhizobium japonicum* are available at SoyKB [57]. A gene was considered expressed if the RPKM value was greater than or equal to two in an expression atlas (modified criteria

from [58]). The RPKM normalized read count data of expressed genes was log_2-transformed and displayed in the form of heatmaps for each subfamily. The heatmap was generated in R [59] using the heatmap.2 function available in the gplots CRAN library. Genes in the heatmaps were ordered for consistency with the phylogeny.

4.3.4 Plant material and stress experiment

The seeds of *G. max* cv. Williams 82 were germinated on moist germination paper and were allowed to grow until the v1 stage (first trifoliolate stage) in a growth chamber maintained at 77 F and 60% humidity throughout the experiment. The temperature and humidity were continuously monitored and maintained in the growth chamber. The salt treatment was applied by transferring the seedlings into 100 mM NaCl solution. For the dehydration treatment, plants were removed from the germination paper and left in air under water-limiting conditions to impose dehydration stress. Root tissue was harvested after 0, 1, 6 and 12 hr of stress treatments. Five plants per time point were maintained for each of the stress treatments. In order to verify the gradual imposition of salt stress treatment, electrical conductivity was measured in two fragments of germination paper, after harvesting root tissue from plants exposed to salt stress at each of the time points (data not shown). Total RNA was isolated using Qiagen RNeasy® Plant mini kit from three biological replicates per time point per the manufacturer's protocol. The RNA samples were treated with Ambion® TURBO DNA-free™ DNase to get rid of any DNA contamination in the RNA samples. The RNA samples were inspected for their quality and quantity using NanoDrop® spectrophotometer and Qubit® fluorometer.

4.3.5 Sequencing, data processing, gene expression analysis and annotation under stress conditions

Total RNA from 21 samples that includes three control samples (0 hr), and three biological replicates for each of the three time points 1, 6 and 12 hr under dehydration and salt stress was sent to the National Center for Genome Resources (Santa Fe, NM, USA) for sequencing on Illumina® HiSeq 2000. Seven randomly chosen samples were multiplexed in each lane and three lanes of HiSeq 2000 were utilized to generate single-end short-reads of 1x50 bp lengths. The reads were aligned with GSNAP [60] using default settings with a maximum of 4 mis-matches allowed against the *Glycine max* genome assembly and annotation v1.01 from Phytozome (JGI Glyma1.0 gene calls). The uniquely mapped reads that mapped to a single location in the genome were analyzed for differential gene expression between the control and treatment samples using the R package DESeq v1.7.10 [61]. A gene was considered to be DE if it satisfied the following three stringent filtering criteria: (1) *P*-value adjusted for multiple testing correction using Benjamini and Hochberg method [62] to be less than 0.05, (2) two fold or greater fold change, (3) residual variance quotients of both the control and treatment samples of less than 20. The residual variance criterion was used to filter genes that have significant variation between replicates, per recommendations in the DESeq manual (Released April 20, 2011). The raw and the normalized read counts, and the sequence data has been deposited in NCBI's Gene Expression Omnibus [63,64] and are accessible through the GEO series accession number GSE57252.

The DE genes were annotated using the top *Arabidopsis* hit, and the corresponding gene ontology (GO) biological process and molecular function terms were inferred [65]. The DE genes under dehydration and salt stress were then screened separately for overrepresented

GO terms against all soybean genes using Fisher's exact test [66] and Bonferroni [67] corrected significance value of less than 0.05. The overrepresented GO terms were enriched at the second level using BLAST2GO v.2.7.1 [68] and a reduced representation of enriched GO terms was obtained. The DE genes were also annotated using the SoyDB [69,70] transcription factor (TF) database, and Fishers's exact test followed by Bonferroni correction was utilized to determine the overrepresented TF classes under each of the stress conditions.

4.3.6 Screening of HD-Zip gene promoters for conserved motifs of transcription factor binding sites (TFBSs)

For the purpose of this study, the one kilobase (kb) region upstream of the annotated transcription start site for each gene was evaluated for promoter motifs. Promoter sequences were retrieved using custom Perl scripts for all gene models in the soybean genome. Promoter sequences that were either less than one kb or included two or more Ns were excluded from the analysis. The program Clover [71] was used to scan through a database of known motifs in TRANSFAC® v. 2010.4 [72]. Promoters of HD-Zip genes belonging to each subfamily were scanned separately for enriched motifs against a background of all soybean gene promoters, with a *P*-value threshold of 0.05 and an individual motif hit score of greater than or equal to 6. Similarly, promoters of genes that were DE in at least one of the three time points under dehydration and salt stress were screened to identify overrepresented motifs under each of the stress treatments. The overrepresented motifs were filtered to include only plant motifs. A comparison was made between motifs that were overrepresented in the promoters of HD-Zip genes belonging to each of the subfamilies and dehydration and salt stress treatments.

4.4 Results

4.4.1 Classification of HD-Zip genes using phylogenetic analysis

A BLASTP search with the Arabidopsis HD-Zip genes against soybean, M. truncatula, grape and rice, followed by reconstruction of the phylogeny, clustered the sequences into four previously defined HD-Zip subfamilies (I to IV). HMMs for each subfamily were used to determine subfamily membership and refine alignments. The outlier sequences excluded from the preliminary tree (see methods for details) included six sequences that belonged to the HD-Zip IV subfamily and these were included in the final phylogenetic trees of the four subfamilies (Figures 1, 2, 3 and 4).

Based on the species clustering patterns and the number of copies of genes belonging to each species, we identified four old angiosperm clades in HD-Zip II, III and IV, and five clades in HD-Zip I (Figures 1, 2, 3 and 4). The topology of most of the angiosperm clades is generally consistent with the species tree. Typically the two legume species (soybean and *M. truncatula*) form a clade, with *Arabidopsis*, grape, and rice each as increasingly distant outgroups from the legume sequences in the clade. The number of copies of genes of each species in each angiosperm clade reflects the number of WGD events the species has undergone. For instance, four of the five angiosperm clades in HD-Zip I phylogeny included exactly three grape sequences – likely the result of the "gamma" triplication event that occurred around 117 Mya [29,30], and the angiosperm clade A1 in the HD-Zip I phylogeny contains nine of the 12 possible soybean sequences - possibly the result of "gamma" triplication event (~117 Mya), and the legume- (~59 Mya) and *Glycine*-specific (~13 Mya) WGD events [30,32]. We identified 101 genes in soybean, 47 in *Arabidopsis*, 33 in grape,

and 41 each in *M. truncatula* and rice (Table 1). The highest gene retention rate (52.7%) among the five species is in HD-Zip IV, whereas the HD-Zip III has the lowest (20.3%) retention rate (Table 1). Although soybean has the highest number of genes, grape and rice have relatively higher retention rate of 64.7% and 60.3% respectively (Table 1). Arabidopsis has the lowest retention rate of 23.0%, whereas soybean and *M. truncatula* have intermediate retention rates of 49.5% and 40.2% respectively (Table 1). The varying rate of retention across the five species reflects the changes in the genomes of each of the species after WGD events. Overall the phylogenetic analysis together with the WGD histories helps clarify our understanding of the evolution of each of the subfamilies.

In the phylogeny generated with sequences from eight species, the eudictos (poplar, cucumber, *Arabidopsis*, grape, soybean and *M. truncatula*) usually clustered together, with the monocots (rice and maize) as an outgroup (Additional file 5: Figure S5, Additional file 6: Figure S6, Additional file 7: Figure S7 and Additional file 8: Figure S8).

4.4.2 Validation of HD-Zip genes using conserved domains, motifs and gene-structures

The HD-Zip I and II sequences contain the Homeobox (PF00046.24) domain and belong to the Homeobox associated leucine zipper family (HALZ; PF02183.13). In addition, the HD-Zip II sequences contain the conserved residues “CPSCE” at the carboxy terminal, and seven of the 24 HD-Zip II sequences contain a HD-ZIP_N (PF04618.7) domain at the N-terminal. The HD-Zip III sequences are highly conserved among all five species along the complete length of the coding sequence (Additional file 3: Figure S3). They contain the Homeobox (PF00046.24), START (PF01852.14) and MEKHLA (PF08670.6) domains. The HD-Zip IV sequences contain the Homeobox (PF00046.24) and the START (PF01852.14) domains. The presence of leucine zipper motif immediately following the homeodomain in

HD-Zip III and IV sequences was confirmed using the sequence logos (Additional file 3: Figure S3, Additional file 4: Figure S4). Exon-intron structures are characteristic for each subfamily (Additional file 9: Figure S9, Additional file 10: Figure S10, Additional file 11: Figure S11 and Additional file 12: Figure S12). The HD-Zip III is particularly conserved, with each gene containing exactly 18 exons. The numbers of exons in genes in the HD-Zip I, II and IV subfamilies are in the ranges 1–5, 3–6, and 8–12. The HD-Zip I and II genes code for smaller proteins, with average peptide length of 265 and 275 amino acids, whereas HD-Zip III and IV genes code for average peptide lengths of 840 and 741 amino acids.

4.4.3 Genomic locations of HD-Zip genes in the soybean genome

The HD-Zip genes are distributed on all 20 chromosomes in the soybean genome, typically in the more gene-dense euchromatic regions near chromosome ends (Figure 5). One HD-Zip II gene (Glyma0041s00350) was found on an unanchored scaffold 41. The HD-Zip genes generally do not occur in clusters or arrays, with only three instances of tandemly duplicated genes.

4.4.4 Genome duplications and expansion of HD-Zip family in the soybean genome

Copy number expansion of the HD-Zip family in the soybean genome has primarily occurred through genome duplication events (Figure 5, Additional file 13: Table S1). Each angiosperm clade in each of the four subfamilies (Figures 1, 2, 3 and 4) contains two to four soybean gene copies that are a result of retention of genes after the legume WGD (~59 Mya) and/or the *Glycine*-specific WGD (~13 Mya). Retention of genes following these WGDs has been high, with retention of 32 of 36 HD-Zip I (88.9%), 20 of 24 HD-Zip II (83.3%), 10 of 11 HD-Zip III (90.9), and 26 of 30 (86.7%) HD-Zip IV genes (Additional file 13: Table S1). There are two tandemly duplicated HD-Zip pairs in subfamily III, and another pair in

subfamily IV. Phylogenetic patterns indicate that the tandemly duplicated genes in subfamily III further duplicated during a WGD event, giving rise to Glyma07g01940 and Glyma07g01950 on chromosome 07 and Glyma08g21610 and Glyma08g21620 on homoeologous chromosome 8. Genes Glyma09g02990 and Glyma09g03000, in HD-Zip IV, are another pair of tandemly duplicated genes. The Glycine WGD event resulted in the homoeologous gene pair Glyma09g03000 and Glyma15gg13950, whereas the homoeologous gene for Glyma09g02990 has evidently either been lost following the WGD – or the Glyma09g02990 and Glyma09g03000 duplication occurred after the Glycine WGD. Overall, 88 of the 101 HD-Zip genes are members of homoeologous gene pairs in the soybean genome.

4.4.5 Expression of HD-Zip genes in 24 conditions including 17 tissues of soybean

The expression of HD-Zip genes was investigated using the *G. max* gene expression atlas reported by Severin et al. [53], and Libault et al. [54,55]. Of the 44 homoeologous gene pairs, 41 show expression in identical tissues (Figures 6, 7, 8, and 9, Additional file 14: Figure S13, Additional file 15: Figure S14, Additional file 16: Figure S15 and Additional file 17: Figure S16). The remaining three show divergent patterns in different tissues between the WGD-derived paralogs (Figures 6 and 9). HD-Zip I gene Glyma06g20230 was expressed in each of the 14 tissues, whereas the homoeolog Glyma04g34340 was expressed in the roots, “pod.shell.10DAF” and “pod.shell.14DAF.” HD-Zip I gene Glyma19g01300 had expression in each of the 14 tissues, but the homoeolog Glyma13g23890 lacked expression in five “seed tissues” (10 DAF, 14 DAF, 21 DAF, 25 DAF and 28 DAF). HD-Zip IV gene Glyma11g00570 was expressed only in the flower, whereas the homoeolog Glyma01g45070 was expressed in young leaf, flower, “one.cm.pod” and “pod.shell.10DAF.” Similar

divergent gene expression patterns between these homoeologous genes were also noticed in the gene expression atlas reported by Libault et al. [54] (Additional file 14: Figure S13, Additional file 17: Figure S16).

Three HD-Zip I (Glyma12g18720, Glyma06g35050, and Glyma19g44800), and four HD-Zip IV (Glyma08g09440, Glyma15g13950, Glyma09g03000, and Glyma05g33520) genes showed no expression in any of the 14 tissues investigated by Severin et al. [53]. However, we found evidence of expression for Glyma12g18720 - HD-Zip I in the roots subjected to dehydration stress after 12 hr (data generated in this study). Glyma06g35050 - HD-Zip I showed expression in leaf, flower and root tip, whereas Glyma09g03000 and Glyma05g33520 - HD-Zip IV were expressed in green pods and shoot apical meristem respectively in Libault et al. [54]. The remaining three genes had no evidence for expression (Glyma19g44800 - HD-Zip I and Glyma08g09440, Glyma15g13950 - HD-Zip IV) in either of the two atlases. These three genes did not reveal any frame shift mutations when investigated at the sequence level. Hence, might be pseudogenes, or incorrectly predicted gene models – or they may only be expressed in certain tissues or under conditions that have not been sampled in this study.

Based on the mean expression of genes across 14 tissues investigated by Severin et al. [53], HD-Zip I and II genes had relatively higher expression in roots and flowers; HD-Zip III in young leaves, “one cm pod,” and “pod shell 10 days after flowering;” and HD-Zip IV in young leaves and flowers. Similar results were observed using the expression atlas generated by Libault et al. [54], with the exception of highest mean expression of genes belonging to each of the subfamilies was noticed in shoot apical meristem. Overall, HD-Zip genes had expression in each of the 17 tissues.

The screening of HD-Zip gene expression using mock-inoculated and *B. japonicum*-infected root hair cells at different time points highlighted HD-Zip genes with more than two fold expression differences between control and treatment samples (Additional file 18: Figure S17, Additional file 19: Figure S18, Additional file 20: Figure S19 and Additional file 21: Figure S20). More than 50% of the genes belonging to the HD-Zip III showed greater than two-fold difference between control and treatment samples at least at one time point (Additional file 20: Figure S19).

4.4.6 Expression of HD-Zip genes under dehydration and salt stress using RNA-Seq

To identify HD-Zip family members responsive to abiotic stress, we used an RNA-seq approach. Twenty-one samples were analyzed by RNA-seq including three control samples (0 hr), and three biological replicates for each of the three time points 1, 6 and 12 hr under dehydration and salt stress. The total number of reads generated in the RNA-Seq experiment from sequencing of 21 sample libraries was 238.8 million, of which 181.2 million (75.9%) uniquely mapped to a single location in the soybean genome (Table 2).

We identified 4,389 and 8,077 genes to be DE in at least one of the three time points (1, 6 or 12 hr) under dehydration and salt stresses respectively (Additional file 22: Table S2, Additional file 23: Table S3, Additional file 24: Table S4, Additional file 25: Table S5, Additional file 26: Table S6 and Additional file 27: Table S7) (see Methods for the filtering criteria). Salt stress resulted in mostly upregulation of genes, whereas dehydration stress caused downregulation of genes (Additional file 28: Table S8). The number of genes discarded from the differential expression analysis due to significant amount of variation between the replicates under dehydration and salt stress at a given time point ranged from 119 to 220 (Additional file 28: Table S8). The raw and DESeq-normalized expression values

for each gene model under both dehydration and salt stress at 1, 6 and 12 hr are provided in Additional file 29: Table S9 and Additional file 30: Table S10 respectively.

Six genes were DE at least at one of the three time points under dehydration stress (Figure 10): five in HD-Zip I, and one in HD-Zip II. Two genes were upregulated and the remaining four were downregulated under dehydration stress. Glyma01g04890 was significantly DE at two different time points. Three of the five DE HD-Zip I genes (Glyma17g10490, Glyma06g20230 and Glyma05g01390) belong to the angiosperm clade A5 (Figure 1), and were a result of the early-legume WGD and the recent *Glycine* WGD.

We found sixteen genes DE at one of the three time points under salt stress (Figure 11): seven in HD-Zip I, four in HD-Zip II, one in HD-Zip III, and four in HD-Zip IV. Nine genes were upregulated and the remaining seven genes were downregulated under salt stress. Five of the 16 genes were significantly DE at two time points (HD-Zip I: Glyma01g04890, Glyma07g05800; HD-Zip II: Glyma15g18320, Glyma13g00310; HD-Zip IV: Glyma13g43350). Four of the seven DE HD-Zip I genes were two homoeologous gene pairs (Glyma07g05800/Glyma16g02390; Glyma01g38390/Glyma11g06940). One of the pairs is a member of angiosperm clade A3, and the other belongs to angiosperm clade A1 (Figure 1). One pair each from the HD-Zip II and HD-Zip IV DE genes (Glyma15g18320/Glyma13g00310 and Glyma13g43350/Glyma07g02220, respectively) resulted from the early-legume WGD. The HD-Zip IV gene Glyma13g38430 was not expressed under the control condition (0 hr time point), but was upregulated after 12 hr under salt stress.

The two HD-Zip I genes, Glyma01g04890 and Glyma16g02390, were DE under both dehydration and salt stress. Glyma01g04890 was upregulated at the 6 hr and 12 hr time

points under both stress treatments, whereas Glyma16g02390 was downregulated at the 6 hr time point under dehydration stress, and upregulated at the 12 hr time point under salt stress. In summary, 20 of the 101 HD-Zip genes in soybean were DE under either dehydration or salt stress, at least at one time point. Eleven of these 20 genes shared a common ancestor either before the early-legume or the *Glycine* WGDs, implying conservation of gene functions following these genome duplications.

4.4.7 Annotation of differentially expressed genes under dehydration and salt stress

In order to help evaluate and confirm results from the application of dehydration and salt stress treatments, GO and TF enrichment analysis were performed on the DE genes. Under dehydration stress, 28 "biological process" and 15 "molecular function" terms were significantly (corrected P <0.05) overrepresented, whereas 41 "biological process" and 27 "molecular function" terms were significantly (corrected P <0.05) overrepresented under salt stress (Additional file 31: Table S11). The enriched biological processes and molecular functions include terms such as - "GO:0009414 - response to water deprivation," "GO:0015250 - water channel activity," and "GO:0009651 - response to salt stress," consistent with the experimental treatments (dehydration and salt stress). At the second level of GO analysis, the biological process category "response to stimulus" was the most prevalent one under both stress treatments, followed by "cellular process" and "metabolic process" (Figure 12A), while in the molecular function category, "catalytic activity" and "binding" were highly represented (Figure 12B).

We identified 503 and 862 TFs among the DE genes under dehydration and salt stress treatments respectively (Additional file 32: Table S12). These TFs corresponded to 35 and 47 TF classes under dehydration and salt stress. Using the enrichment analysis, we identified

four TF classes, “WRKY,” “AP2-EREBP,” “ZIM” and “C2C2 (Zn) CO-like” to be significantly (corrected P <0.05) overrepresented under both stress treatments, whereas the TF class “NAC” was overrepresented only under salt stress (Table 3).

4.4.8 Promoter analysis

The enrichment analysis performed with the Clover program [71] and the TRANSFAC database [72] on the promoters of HD-Zip genes identified four different transcription factor binding sites (TFBSs) overrepresented in the promoters of HD-Zip I genes, and at least 9 different TFBSs in HD-Zip II to IV genes (Table 4). The genes belonging to the same subfamily had a diverse profile of TFBSs enriched in the promoters, suggesting the possible role of promoter sequences in functional diversification of the HD-Zip genes of the same subfamilies (Additional file 33: Table S13). The homoeologous genes in all subfamilies had reasonably different TFBSs enriched in their promoters, suggesting specific regulation of homoeologous genes under particular conditions (Additional file 33: Table S13).

There are 14 TFBSs overrepresented in the promoters of HD-Zip genes as well as promoters of DE genes under dehydration stress. Similarly nine TFBSs are overrepresented in the promoters of HD-Zip genes and the promoters of DE genes under salt stress (Table 4, Additional file 34: Table S14). These TF classes are potential candidates that may influence both HD-Zip genes as well genes involved in dehydration and salt stress responses.

The TFBSs “Dof3” and “PBF” are overrepresented in more than 90% of the HD-Zip I and IV genes respectively, and “Alfin1” is overrepresented in more than 90% of HD-Zip II and III genes (Table 4). Hence, these transcription factors probably play an important role in regulating certain HD-Zip genes.

Finally, all but three TF classes corresponding to enriched TFBSs in the promoters of HD-Zip genes contain DE genes under dehydration and salt stress (Table 4). This observation is consistent with HD-Zip genes playing important roles under dehydration and salt stress-responses.

4.5 Discussion

4.5.1 Identification and phylogenetic analysis of HD-Zip genes

In this study we have identified and characterized 101 HD-Zip genes in the soybean genome. Recently, 88 HD-Zip genes have been described in soybean [12]. Chen et al. [12] used BLASTP to identify 100 putative HD-Zip transcription factors. SMART and PFAM analyses requiring both an HD and LZ domain were used to refine the number of HD-Zip genes to 88. Similarly, we initiated our study using BLASTP of Arabidopsis HD-Zip genes against the proteomes of soybean, *M. truncatula*, rice and grape. We then used phylogenetic analyses coupled with HMM searches, domain analyses, and known evolutionary relationships among the five species, to identify more diverse members of the HD-Zip family in each of these species. Using this approach, we were able to identify 13 additional novel HD-Zip genes in soybean and identify the HD-Zips in *M. truncatula* and grape, which had previously been unreported. Not surprisingly, our approach had the biggest impact on the largely uncharacterized HD-Zip IV genes. While Chen et al. [12] reported 19 genes in HD-Zip IV, we have found 30 genes. These genes may have novel biological functions.

By including multiple species in our search for HD-Zip genes, we also improved the classification of the different family members in soybean and other species. The clustering of

Arabidopsis genes in the HD-Zip subfamilies was consistent with the results of Ariel et al. [1]. The HD-Zip I and II subfamilies can be classified into nine (α, β1, β2, γ, δ, ε, ϕ1, ϕ2 and ζ) and four (α, β, γ and δ) clades respectively that have been previously described in studies on *Arabidopsis*, rice and maize [4,7,8] (data not shown). Although the results in our study are consistent with the later classification, we suggest that the later strategy be used with discretion. One instance where it can lead to conflicting results is that the ζ clade has been described as monocot-specific clade in all previous studies [4,7,8], but this clade clearly contained dicot sequences as a part of an old angiosperm clade in our study. One potential reason for this conflict is that the previous studies included only *Arabidopsis* [4,8] or *Arabidopsis* and *C. plantagineum* [7] as the dicot species. Sampling of additional dicot sequences of soybean, *M. truncatula* and grape in this study provided a clearer picture of the taxonomic contexts of the HD-Zip gene family.

We identified five ancient angiosperm clades in HD-Zip I, and four in the HD-Zip II, III and IV subfamilies. The presence of these multiple angiosperm clades in each subfamily is consistent with the recent discovery of two ancient WGD events, one occurring at the base of the angiosperm lineage (ancient angiosperm WGD) and the other before the angiosperm-gymnosperm split (ancestral seed plant WGD) [31]. Early diversification driven by multiple early-plant WGDs is also consistent with a previous study of the evolution of HD-Zip III subfamily in land plants [73]. The presence of five angiosperm clades in HD-Zip I (rather than the four that would be expected from two early WGDs) is intriguing and needs further investigation in the context of synteny analysis and inclusion of additional species in the phylogeny.

A phylogeny with eight species, including published HD-Zip sequences from maize, poplar and cucumber, was largely congruent with the phylogeny generated using five species. These phylogenetic relationships will help identify orthologous genes, and accelerate functional characterizations studies.

4.5.2 Conserved domains and gene structures for validation of HD-Zip genes

PFAM and sequence logos identify highly conserved domains and motifs in the HD-Zip gene family. These have been reported in previous studies [1,2,4,12], but we note two exceptions: *Arabidopsis* HD-Zip I gene AT1G27050 had an additional "RRM_1" (RNA recognition motif), and *Medicago* HD-Zip IV gene Mt.ctg127898_1 had two START domains. Overall, the highly conserved domains and motifs are the signatures of the HD-Zip gene family and can be utilized to validate genes identified using several approaches.

Exon-intron structures are generally well conserved in each HD-Zip subfamily, particularly within each angiosperm clade. The HD-Zip III gene-structures were remarkably conserved, with each of the soybean genes having precisely 18 exons. Considering HD-Zip III gene structures reported in other species, all genes in poplar had 18 exons [9], and 4 of the 5 maize genes had 18 exons [8], but in rice only one of the four genes had 18 exons [7]. The generally well-conserved exon structure in HD-Zip III genes across different species highlights the possibility of conserved gene function and strict regulation of these genes. In a recent study involving identification of genes that are potential targets of miRNA in developing soybean seeds, all HD-Zip III genes were found to be targets of miRNA 166 [74]. Prigge and Clark [73], and Floyd and Bowman [75] have previously suggested that HD-Zip III sequences across all land plants produce transcripts that could be targeted by miRNA165 and miRNA166. DeRocher and Nguyen [76] overexpressed *Arabidopsis* HD-Zip III gene

REVOLUTA in soybean embryo, leading to seed yield increase with no change in the seed composition. In short, the HD-Zip III genes appear to be both highly conserved and under intricate transcriptional regulation.

4.5.3 Expansion of HD-Zip gene family

The 101 HD-Zip genes in soybean is the highest number reported so far in any angiosperm species, comparing with 48 in *Arabidopsis* [2], 55 in maize [8], 47 in rice [2], and 63 in poplar [9]. The HD-Zip genes in soybean have expanded during the early-legume WGD event (~59 Mya), and the *Glycine* WGD event (~10 Mya), with high retention of paralogs. Expansion of the HD-Zip gene family due to WGDs has been previously reported in other species. The *Arabidopsis*, rice, maize and poplar have at least 75%, 50%, 62% and 81% homoeologous gene pairs respectively [5,6,8,9,77-79]. However, in cucumber, a species that lacks WGD events since eudicot radiation, there are no homoeologous gene pairs among HD-Zip I and IV (the two subfamilies described in cucumber) [10,11]. These results imply that the HD-Zip gene family has expanded in a species-specific manner, with copy number generally depending on WGD events and high retention rates after duplications.

Gene families can be broadly categorized as having high rates of retention of segmental (WGD-derived) duplicates and low generation or retention of tandem duplicates – or vice versa (low segmental retention, high tandem generation and retention) [80]. The low-tandem/high-segmental duplication class of gene families has been reported to comprise highly conserved, housekeeping, and key regulatory gene families [80] – for example, transcription factor families such as heat shock and WRKY, housekeeping families such as mitochondrial carrier proteins [81,82], and the proteasome 20S subunit family [83,84]. Clearly, the HD-Zip superfamily falls in the “high segmental, low tandem” category, with

only three tandem duplication events in the HD-Zip genes in soybean. The expansion and retention of the HD-Zip family during segmental duplication events will have consequences for functional characterization studies, due to the possibility of genetic redundancy in duplicated genes.

4.5.4 Gene expression patterns of HD-Zip genes in 24 conditions, including 17 tissues

The *G. max* expression atlas [53] was initially utilized for investigating gene expression patterns of HD-Zip genes in 14 tissues of soybean. The average expression values across 14 tissues for each subfamily was highly variable, and there were genes with extremely high expression relative to the average expression across tissues in each of the subfamily. Investigating gene expression patterns separately for each subfamily on a $\log_2$-transformed scale helped identify gene expression patterns that were unreported in Chen et al. [12]. Chen et al. [12] displayed expression of all four subfamilies on a single scale using average linkage clustering method. In addition we utilized two additional gene expression atlases developed by Libault et al. [54,55], which allowed investigation of HD-Zip genes in three additional tissues, and seven different conditions.

All but three homoeologous gene pairs show consistent expression in the same tissues between the WGD-derived paralogs, suggesting retention of HD-Zip gene functions after genome duplications. The genome duplication events provide raw materials for new gene functions. The duplicated gene can evolve to have a new function (neofunctionalization) [85] or can acquire new deleterious mutations and become a pseudogene (pseudogneization); or both the ancestral and the newly formed gene can undergo reduction in their levels and patterns of activity, such that jointly their function matches with that of the ancestral gene (subfunctionalization) [86].

4.5.5 RNA-Seq based expression profiling of soybean genes during dehydration and salt stress

RNA-Seq analysis was utilized to investigate genes involved in dehydration and salt stress. The expression of all soybean genes including the 101 HD-Zip genes identified in this study was studied in the roots of soybean cv. Williams 82 at V1 stage, at four different time points, and under dehydration and salt stress. The evaluation of plants at the V1 stage may assist in identification of candidate genes involved in initiation of dehydration and salt stress. Recently, Chen et al. [12] reported the influence of drought and salinity stress on HD-Zip genes using publicly available microarray data sets available at National Center for Biotechnology Information under accession numbers GSE41125 and GSE40627. The microarray datasets facilitated investigation of the expression of 55 of the 88 HD-Zip genes identified in their study. The microarray dataset GSE40627 reports expression of genes in the leaves under drought stress imposed at late developmental stages (V6 and R2), whereas the dataset GSE41125 describes expression of genes in 14 d seedlings utilizing pooled RNA samples from 0, 3, 6, 12 and 24 hr of mock and salinity stressed plants. Thus, in the current study, the utilization of root tissue at the V1 stage, and investigation of gene expression separately at each of the four time points 0, 1, 6 and 12 hr provided clearly different and more precise insight into genes that are involved in dehydration and salt stress.

We identified 4,389 and 8,077 genes to be DE in the roots of soybean cv. Williams 82 at the V1 stage at least at one of the three time points (1, 6 or 12 hr) under dehydration and salt stress respectively. Partial validation of DE genes for their role in abiotic stress responses was obtained by performing GO and TF enrichment analysis. The highly represented biological process GO categories, “response to stimulus,” “cellular process,” and “metabolic

process" as well as the molecular function categories, "catalytic activity" and "binding," are generally found to be enriched during abiotic stress responses [87-89]. Similarly, the four TF classes WRKY [90-92], AP2-EREBP [93-95], ZIM [96-98] and C2C2 (Zn) CO-like [99-101] (all enriched under both dehydration and salt stresses), and NAC [102-104] (overrepresented under salt stress) are major TFs that have previously been shown to play critical role in stress responses, and are consistent with results reported in this study.

4.5.6 Expression profiling of HD-Zip genes under dehydration stress

RNA-Seq analysis identified 20 HD-Zip genes DE in the roots of soybean cv. "Williams 82," under dehydration and salt stress. The role of HD-Zip genes in regulation of developmental adaptation under different environmental stress conditions has been previously established in *Arabidopsis*, *Medicago*, rice, sunflower, maize, cucumber, and poplar [4,7-11,19,105-108].

All six genes identified as DE in the roots under dehydration stress in this study, were also, DE under drought stress in leaves [12]. Four of the five DE HD-Zip I genes belong to the angiosperm clade A5. This clade contains genes such as *CPHB-5* from *C. plantagineum*, and *Zmhdz1*, *−2*, *−3* from maize, that have previously been shown to have a role in water-stress response [4,8,20]. Chen et al. [109] showed Glyma06g20230 DE in this study was DE under dehydration stress, in the roots of drought-tolerant soybean genotype, "Jindou21."

HD-Zip I gene Glyma16g02390 that is DE under dehydration stress belongs to the angiosperm clade A3. Genes in this clade have been extensively characterized for their role in water-stress responses in other species. For example, the *Arabidopsis* ATHB7 and ATHB12 genes have been shown to reduce plant growth under water-deficit condition [13,16,17]. The sunflower *HaHB4* gene is strongly induced by water deficit stress [14], and

when over-expressed in *Arabidopsis* the plants exhibit increased survival by a process that inhibits-drought related senescence [18,19]. The *N. attenuata NaHD20* gene is induced in roots under water-deficit conditions [15]. The rice *Oshox6*, *22* and *24* genes are involved in drought-responsiveness [7]. Hence, we hypothesize that the soybean gene Glyma16g02390 may have a role under water-deficit stress response and is a potential candidate for functional characterization.

HD-Zip II gene Glyma08g15780 that is DE under dehydration stress is an ortholog of rice genes *Oshox11* and *Oshox27*, which have also been demonstrated to be involved in drought-response [7].

In summary, the HD-Zip I and II genes show differential expression patterns under dehydration stress that are consistent with the water-deficit stress response functions of orthologous genes previously identified in studies of water stress. These results support that HD-Zip I and II genes may generally have a role, conserved across many angiosperm species, in mediating water-stress responses; and that these genes may be viable targets for developing more drought-tolerant soybean cultivars.

4.5.7 Expression profiling of HD-Zip genes under salt stress

A subset of HD-Zip genes, from each of the four subfamilies, responded to salt (100 mM NaCl) stress in the roots, in at least one of the three time points. Six of the 16 genes (Glyma01g04890, Glyma07g05800, Glyma16g02390, Glyma13g05270, Glyma15g18320, Glyma03g30200) DE under salt stress have been recently shown to respond to salt stress, in 14 d old seedlings of soybean plant, in a microarray experiment [12].

The HD-Zip I gene Glyma13g05270 was downregulated under salt stress, which is similar to the expression of its Arabidopsis orthologs, *ATHB3* and *ATHB20*, which are

similarly downregulated under salt stress [5]. The homoeologous genes Glyma01g38390 and Glyma11g06940 were upregulated after 12 hr of salt stress, comparable to the *Arabidopsis* orthologs, *ATHB20*, *ATHB50* and *ATHB53*, which are upregulated more than two-fold under salt stress [5].

Two of the four DE HD-Zip IV genes, Glyma13g43350 and Glyma13g38430, had nearly zero expression under control conditions, but were upregulated under salt stress, suggesting a possible role in root development under stress conditions. Glyma1343350 and Glyma07g02220 are orthologs of the *Arabidopsis* gene GLABRA2, which has been functionally characterized and shown to regulate root hair development, and cell specification of root epidermis in salt stressed plants [78,110-112].

The two homoeologous HD-Zip I genes (Glyma07g05800 and Glyma16g02390) upregulated under salt stress belong to the angiosperm clade "A3." This clade contains the functionally characterized *Medicago* gene MtHB1 (Medtr8g026960). *MtHB1* is induced in the roots under ABA and salt stress, and regulates lateral root emergence in *Medicago* [26]. The reduction of lateral root emergence by MtHB1, under salt stress, is a mechanism to minimize the exposure of plant roots to excess salt in the soil.

The HD-Zip I gene Glyma01g04890 was upregulated at 6 and 12 hr time points under both salt and dehydration stress. This gene was also upregulated under both drought and salt stress in the leaves and seedlings, respectively, in two microarray experiments [12]. A BLASTP search with Glyma01g04890 protein sequence against the patent database [113,114] found a match (E-value =0; Similarity >99.4%; Coverage =100%) with sequences in five "patent applications" (US_2012_0278947_A1; US_2012_0096584_A1;

US_2007_0277269_A1; US_2012_0005773_A1; US_2009_0144847_A1) that described the role of this sequence in improving plant performance under abiotic stress.

4.5.8 Functional diversity and regulation of HD-Zip genes

The presence of highly diverse TFBSs enriched in the promoters of HD-Zip genes provides evidence for functional diversity. Previous studies have mainly focused on HD-Zip target-sequences, and regulatory regions adjacent to the DNA-binding domain of HD-Zip genes. All experimentally tested HD-Zip I genes have been shown to bind specifically, and with high affinity to target-sequences comprising of the same pseudopalindromic sequence CAAT(A/T)ATTG, under *in vitro* conditions [115-117]. Arce et al. [118] reported the presence of activation domain, sumoylation, and phosphorylation sites in the carboxy-terminal regions, and some putative regulatory regions in the amino-terminal regions, as being responsible for the functional diversity of HD-Zip I genes.

The “Dof3” and “PBF” TFBSs are enriched in more than 90% of HD-Zip I and IV gene promoters respectively. The “Dof” TFs like HD-Zip are plant-specific TFs and are involved in several process, for example stress-responses [119-121], phytochrome signaling [122], light-responses [123,124], responses to plant hormones including auxin [125,126] and gibberellin [127,128], and seed germination [129,130]. PBF also known as whirly family are known to regulate plant defense gene expression [131].

The TFBS “Alfin1” is overrepresented in more than 90% of HD-Zip II and III gene promoters. “Alfin1” TFs are shown to contribute toward salt tolerance in plants [132,133].

Finally, the presence of highly diverse TFBSs enriched in the promoters of HD-Zip genes, both within and across subfamilies, suggests the complex integration of HD-Zip genes

in various signal-transduction pathways, with a potential source for functional diversity of these highly conserved HD-Zip genes.

4.6 Conclusions

In this study we have described the soybean HD-Zip gene superfamily. Evolutionary histories, interpreted in the context of whole genome duplication events and analysis of gene structures, provide additional verification for the classification of the soybean HD-Zip genes. The HD-Zip genes in the soybean genome were preferentially retained after the legume-specific and/or *Glycine*-specific whole genome duplication events. The RNA-Seq experiment identified candidate genes that may be involved in dehydration and salt stress responses.

Authors' contributions

VB and SBC conceived and planned the project. VB carried out the experiments. VB, SBC, NTW, AKB, ADF and MAG performed data analysis. VB and SBC wrote the manuscript. All authors read and approved the final manuscript.

4.7 Acknowledgments

The authors are thankful to Dr. Randy C. Shoemaker for providing laboratory facilities, and to Rebecca Nolan for her invaluable support during laboratory experiments. This work was supported by USDA-ARS project funds to Steven B. Cannon.

4.8 References

1. Ariel FD, Manavella PA, Dezar CA, Chan RL: The true story of the HD-Zip family. *Trends Plant Sci* 2007, 12:419–426.
2. Mukherjee K, Brocchieri L, Burglin TR: A comprehensive classification and evolutionary analysis of plant homeobox genes. *Mol Biol Evol* 2009, 26:2775–2794.
3. Zalewski CS, Floyd SK, Furumizu C, Sakakibara K, Stevenson DW, Bowman JL: Evolution of the class IV HD-zip gene family in streptophytes. *Mol Biol Evol* 2013, 30:2347–2365.
4. Harris JC, Hrmova M, Lopato S, Langridge P: Modulation of plant growth by HD-Zip class I and II transcription factors in response to environmental stimuli. *New Phytol* 2011, 190:823–837.
5. Henriksson E, Olsson AS, Johannesson H, Johansson H, Hanson J, Engstrom P, Soderman E: Homeodomain leucine zipper class I genes in Arabidopsis: expression patterns and phylogenetic relationships. *Plant Physiol* 2005, 139:509–518.
6. Ciarbelli AR, Ciolfi A, Salvucci S, Ruzza V, Possenti M, Carabelli M, Fruscalzo A, Sessa G, Morelli G, Ruberti I: The Arabidopsis homeodomain-leucine zipper II gene family: diversity and redundancy. *Plant Mol Biol* 2008, 68:465–478.
7. Agalou A, Purwantomo S, Overnas E, Johannesson H, Zhu X, Estiati A, de Kam RJ, Engstrom P, Slamet-Loedin IH, Zhu Z, Wang M, Xiong L, Meijer AH, Ouwerkerk PB: A genome-wide survey of HD-Zip genes in rice and analysis of drought-responsive family members. *Plant Mol Biol* 2008, 66:87-103.
8. Zhao Y, Zhou Y, Jiang H, Li X, Gan D, Peng X, Zhu S, Cheng B: Systematic analysis of sequences and expression patterns of drought-responsive members of the HD-Zip gene family in maize. *PLoS One* 2011, 6:e28488.
9. Hu R, Chi X, Chai G, Kong Y, He G, Wang X, Shi D, Zhang D, Zhou G: Genome-wide identification, evolutionary expansion, and expression profile of homeodomain-leucine zipper gene family in poplar (Populus trichocarpa). *PLoS One* 2012, 7:e31149.
10. Liu W, Fu R, Li Q, Li J, Wang L, Ren Z: Genome-wide identification and expression profile of homeodomain-leucine zipper class I gene family in cucumis sativus. *Gene* 2013, 531:279–287.
11. Fu R, Liu W, Li Q, Li J, Wang L, Ren Z: Comprehensive analysis of the homeodomain-leucine zipper IV transcription factor family in Cucumis sativus. *Genome* 2013, 56:395–405.
12. Chen X, Chen Z, Zhao H, Zhao Y, Cheng B, Xiang Y: Genome-wide analysis of soybean HD-zip gene family and expression profiling under salinity and drought treatments. *PLoS One* 2014, 9:e87156.

13. Olsson A, Engstrom P, Soderman E: The homeobox genes ATHB12 and ATHB7 encode potential regulators of growth in response to water deficit in Arabidopsis. *Plant Mol Biol* 2004, 55:663–677.

14. Gago GM, Almoguera C, Jordano J, Gonzalez DH, Chan RL: Hahb-4, a homeobox-leucine zipper gene potentially involved in abscisic acid-dependent responses to water stress in sunflower. *Plant Cell Environ* 2002, 25:633–640.

15. Re DA, Dezar CA, Chan RL, Baldwin IT, Bonaventure G: Nicotiana attenuata NaHD20 plays a role in leaf ABA accumulation during water stress, benzylacetone emission from flowers, and the timing of bolting and flower transitions. *J Exp Bot* 2011, 62:155–166.

16. Hjellstrom M, Olsson ASB, Engstrom P, Soderman EM: Constitutive expression of the water deficit-inducible homeobox gene ATHB7 in transgenic Arabidopsis causes a suppression of stem elongation growth. *Plant Cell Environ* 2003, 26:1127–1136.

17. Son O, Hur YS, Kim YK, Lee HJ, Kim S, Kim MR, Nam KH, Lee MS, Kim BY, Park J, Lee SC, Hanada A, Yamaguchi S, Lee IJ, Kim SK, Yun DJ, Soderman E, Cheon CI: ATHB12, an ABA-inducible homeodomain-leucine zipper (HD-Zip) protein of Arabidopsis, negatively regulates the growth of the inflorescence stem by decreasing the expression of a gibberellin 20-oxidase gene. *Plant Cell Physiol* 2010, 51:1537-1547.

18. Dezar CA, Gago GM, González DH, Chan RL: Hahb-4, a sunflower homeobox-leucine zipper gene, is a developmental regulator and confers drought tolerance to Arabidopsis thaliana plants. *Transgenic Res* 2005, 14:429–440.

19. Manavella PA, Arce AL, Dezar CA, Bitton F, Renou JP, Crespi M, Chan RL: Cross-talk between ethylene and drought signalling pathways is mediated by the sunflower Hahb-4 transcription factor. *Plant J* 2006, 48:125–137.

20. Deng X, Phillips J, Meijer A, Salamini F, Bartels D: Characterization of five novel dehydration-responsive homeodomain leucine zipper genes from the resurrection plant Craterostigma plantagineum. *Plant Mol Biol* 2002, 49:601–610.

21. Johannesson H, Wang Y, Hanson J, Engstrom P: The Arabidopsis thaliana homeobox gene ATHB5 is a potential regulator of abscisic acid responsiveness in developing seedlings. *Plant Mol Biol* 2003, 51:719–729.

22. Soderman E, Hjellstrom M, Fahleson J, Engstrom P: The HD-Zip gene ATHB6 in Arabidopsis is expressed in developing leaves, roots and carpels and up-regulated by water deficit conditions. *Plant Mol Biol* 1999, 40:1073–1083.

23. Himmelbach A, Hoffmann T, Leube M, Hohener B, Grill E: Homeodomain protein ATHB6 is a target of the protein phosphatase ABI1 and regulates hormone responses in Arabidopsis. *EMBO J* 2002, 21:3029–3038.

24. Leung J, Merlot S, Giraudat J: The Arabidopsis ABSCISIC ACID-INSENSITIVE2 (ABI2) and ABI1 genes encode homologous protein phosphatases 2C involved in abscisic acid signal transduction. *Plant Cell* 1997, 9:759–771.

25. Deng X, Phillips J, Brautigam A, Engstrom P, Johannesson H, Ouwerkerk PF, Ruberti I, Salinas J, Vera P, Iannacone R, Meijer A, Bartels D: A homeodomain leucine zipper gene from *craterostigma plantagineum* regulates abscisic acid responsive gene expression and physiological responses. *Plant Mol Biol* 2006, 61:469-489.

26. Ariel F, Diet A, Verdenaud M, Gruber V, Frugier F, Chan R, Crespi M: Environmental regulation of lateral root emergence in Medicago truncatula requires the HD-Zip I transcription factor HB1. *Plant Cell* 2010, 22:2171–2183.

27. Huang D, Wu W, Abrams SR, Cutler AJ: The relationship of drought-related gene expression in Arabidopsis thaliana to hormonal and environmental factors. *J Exp Bot* 2008, 59:2991–3007.

28. Yu L, Chen X, Wang Z, Wang S, Wang Y, Zhu Q, Li S, Xiang C: Arabidopsis enhanced drought tolerance1/HOMEODOMAIN GLABROUS11 confers drought tolerance in transgenic rice without yield penalty. *Plant Physiol* 2013, 162:1378–1391.

29. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, Wu X, Zhang Y, Wang J, Carpenter EJ, Deyholos MK, Kutchan TM, Chanderbali AS, Soltis PS, Stevenson DW, McCombie R, Pires JC, Wong GK, Soltis DE, Depamphilis CW: A genome triplication associated with early diversification of the core eudicots. *Genome Biol* 2012, 13:R3.

30. Proost S, Pattyn P, Gerats T, Van de Peer Y: Journey through the past: 150 million years of plant genome evolution. *Plant J* 2011, 66:58–65.

31. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW: Ancestral polyploidy in seed plants and angiosperms. *Nature* 2011, 473:97-100.

32. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, et al: Genome sequence of the palaeopolyploid soybean. *Nature* 2010, 463:178-183.

33. Young ND, Debelle F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KFX, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KAT, Tang H, Rombauts S, Zhao PX, Zhou P, et al: The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 2011, 480:520-524.

34. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, et al: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007, 449:463-467.

35. The Arabidopsis Information Resource. [http://www.arabidopsis.org/].

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403–410.

37. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, 32:1792–1797.

38. Gouy M, Guindon S, Gascuel O: SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010, 27:221–224.

39. Galtier N, Gouy M, Gautier C: SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 1996, 12:543–548.

40. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, 22:4673–4680.

41. FigTree. [http://tree.bio.ed.ac.uk/software/figtree/].

42. HMMER. [http://hmmer.janelia.org/].

43. WebLogo. [http://weblogo.berkeley.edu/logo.cgi].

44. Guindon S, Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003, 52:696–704.

45. iPlant collaborative. [http://www.iplantcollaborative.org/].

46. Anisimova M, Gascuel O: Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol* 2006, 55:539–552.

47. Pfam. [http://pfam.sanger.ac.uk/search#tabview=tab1].

48. Phytozome. [www.phytozome.net].

49. Bio-graphics. [http://search.cpan.org/~lds/Bio-Graphics/].

50. Cannon EK, Cannon SB: Chromosome visualization tool: a whole genome viewer. *Int J Plant Genomics* 2011, 2011:373875.

51. Phytozome v4.0. [ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v4.0/Gmax/misc_feature/Glyma1_domains/glyma1_syn_par.txt].

52. Wang L, Guo K, Li Y, Tu Y, Hu H, Wang B, Cui X, Peng L: Expression profiling and integrative analysis of the CESA/CSL superfamily in rice. *BMC Plant Biol* 2010, 10:282.

53. Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC: RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol* 2010, 10:160.

54. Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G: An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J* 2010, 63:86–99.

55. Libault M, Farmer A, Brechenmacher L, Drnevich J, Langley RJ, Bilgin DD, Radwan O, Neece DJ, Clough SJ, May GD, Stacey G: Complete transcriptome of the soybean root hair cell, a single-cell model, and its alteration in response to Bradyrhizobium japonicum infection. *Plant Physiol* 2010, 152:541–552.

56. Soybase. [http://soybase.org/soyseq/].

57. soykb. [http://soykb.org/].

58. Woody JL, Severin AJ, Bolon YT, Joseph B, Diers BW, Farmer AD, Weeks N, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC: Gene expression patterns are correlated with genomic and genic structure in soybean. *Genome* 2011, 54:10-18.

59. R: A language and environment for statistical computing. [http://www.R-project.org/].

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

61. Anders S, Huber W: Differential expression analysis for sequence count data. *Genome Biol* 2010, 11:R106.

62. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* 1995, 57:289–300.

63. Edgar R, Domrachev M, Lash AE: Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002, 30:207–210.

64. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A: NCBI GEO: archive for functional genomics data sets - update. *Nucleic Acids Res* 2013, 41:D991-D995.

65. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM: Gene Ontology: tool for the unification of biology. *Nature Genetics* 2000, 25:25.

66. Fisher RA: *The design of experiments.* 8th edition. Edinburg: London Oliver and Boyd; 1966.

67. Bonferroni CE: Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni* 1935, 13–60.

68. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21:3674–3676.

69. Wang Z, Libault M, Joshi T, Valliyodan B, Nguyen H, Xu D, Stacey G, Cheng J: SoyDB: a knowledge database of soybean transcription factors. *BMC Plant Biol* 2010, 10:1–12.

70. SoyDB. [http://casp.rnet.missouri.edu/soydb/].

71. Frith MC, Fu Y, Yu L, Chen JÄ, Hansen U, Weng Z: Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 2004, 32:1372–1381.

72. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 2006, 34:D108-D110.

73. Prigge MJ, Clark SE: Evolution of the class III HD-Zip gene family in land plants. *Evol Dev* 2006, 8:350–361.

74. Song QX, Liu YF, Hu XY, Zhang WK, Ma B, Chen SY, Zhang JS: Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. *BMC Plant Biol* 2011, 11:5.

75. Floyd SK, Bowman JL: Gene regulation: Ancient microRNA target sequences in plants. *Nature* 2004, 428:485–486.

76. Google patents. [http://www.google.com/patents/US8653325].

77. Prigge MJ, Otsuga D, Alonso JM, Ecker JR, Drews GN, Clark SE: Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in Arabidopsis development. *Plant Cell* 2005, 17:61–76.

78. Nakamura M, Katsumata H, Abe M, Yabe N, Komeda Y, Yamamoto KT, Takahashi T: Characterization of the class IV homeodomain-Leucine Zipper gene family in Arabidopsis. *Plant Physiol* 2006, 141:1363–1375.

79. Jain M, Tyagi AK, Khurana JP: Genome-wide identification, classification, evolutionary expansion and expression analyses of homeobox genes in rice. *Febs J* 2008, 275:2845–2861.

80. Cannon SB, Mitra A, Baumgarten A, Young ND, May G: The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. *BMC Plant Biol* 2004, 4:10.

81. Kuan J, Saier MH: The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. *Crit Rev Biochem Mol Biol* 1993, 28:209–233.

82. Borecky J, Maia IG, Arruda P: Mitochondrial uncoupling proteins in mammals and plants. *Biosci Rep* 2001, 21:201–212.

83. Parmentier Y, Bouchez D, Fleck J, Genschik P: The 20S proteasome gene family in Arabidopsis thaliana. *FEBS Lett* 1997, 416:281–285.

84. Vierstra RD: The ubiquitin/26S proteasome pathway, the complex last chapter in the life of many plant proteins. *Trends Plant Sci* 2003, 8:135–142.

85. Hughes AL: The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* 1994, 256:119–124.

86. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 1999, 151:1531–1545.

87. Yao L-M, Wang B, Cheng L-J, Wu T-L: Identification of key drought stress-related genes in the hyacinth bean. *PLoS One* 2013, 8:e58108.

88. Xu J, Yuan Y, Xu Y, Zhang G, Guo X, Wu F, Wang Q, Rong T, Pan G, Cao M, Tang Q, Gao S, Liu Y, Wang J, Lan H, Lu Y: Identification of candidate genes for drought tolerance by whole-genome resequencing in maize. *BMC Plant Biol* 2014, 14:83.

89. Z-h D, Zheng LL, Wang J, Gao Z, Wu SB, Qi Z, Wang YC: Transcriptomic profiling of the salt-stress response in the wild recretohalophyte Reaumuria trigyna. *BMC Genomics* 2013, 14:29.

90. Tang J, Wang F, Wang Z, Huang Z, Xiong A, Hou X: Characterization and co-expression analysis of WRKY orthologs involved in responses to multiple abiotic stresses in Pak-choi (Brassica campestris ssp. chinensis). *BMC Plant Biol* 2013, 13:188.

91. Jing L: *Role of WRKY Transcription Factors in Arabidopsis Development and Stress Responses. PhD thesis.* University of Helsinki, Faculty of Biological and Environmental Sciences, Department of Biosciences; 2014.

92. Chen L, Song Y, Li S, Zhang L, Zou C, Yu D: The role of WRKY transcription factors in plant abiotic stresses. *Biochim Biophys Acta (BBA) Gene Regul Mech* 2012, 1819:120–128.

93. Sharoni AM, Nuruzzaman M, Satoh K, Shimizu T, Kondoh H, Sasaya T, Choi I-R, Omura T, Kikuchi S: Gene structures, classification and expression models of the AP2/EREBP transcription factor family in rice. *Plant Cell Physiol* 2011, 52:344–360.

94. Reddy DS, Mathur PB, Sharma KK: Regulatory role of transcription factors in abiotic stress responses in plants. In *Climate Change and Plant Abiotic Stress Tolerance.* Edited by Tuteja N, Gill SS. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2013:555–588.

95. Kizis D, Lumbreras V, Pagès M: Role of AP2/EREBP transcription factors in gene regulation during abiotic stress. *FEBS Lett* 2001, 498:187–189.

96. Ismail A, Riemann M, Nick P: The jasmonate pathway mediates salt tolerance in grapevines. *J Exp Bot* 2012, 63:2127–2139.

97. Vanholme B, Grunewald W, Bateman A, Kohchi T, Gheysen G: The tify family previously known as ZIM. *Trends Plant Sci* 2007, 12:239–244.

98. Jiang Y, Deyholos M: Comprehensive transcriptional profiling of NaCl-stressed Arabidopsis roots reveals novel classes of responsive genes. *BMC Plant Biol* 2006, 6:25.

99. Yao D, Zhang X, Zhao X, Liu C, Wang C, Zhang Z, Zhang C, Wei Q, Wang Q, Yan H, Li F, Su Z: Transcriptome analysis reveals salt-stress-regulated biological processes and key pathways in roots of cotton (*Gossypium hirsutum L.*). Genomics 2011, 98:47.

100. Mishra S, Shukla A, Upadhyay S, Sanchita, Sharma P, Singh S, Phukan UJ, Meena A, Khan F, Tripathi V, Shukla RK, Shrama A: Identification, occurrence, and validation of DRE and ABRE Cis-Regulatory motifs in the promoter regions of genes of *Arabidopsis thaliana*. J Integr Plant Biol 2014, 56:388-399.

101. Hiz MC, Canher B, Niron H, Turet M: Transcriptome analysis of salt tolerant common bean (Phaseolus vulgaris L.) under saline conditions. *PLoS One* 2014, 9:e92598.

102. Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K: NAC transcription factors in plant abiotic stress responses. *Biochim Biophys Acta (BBA) Gene Regul Mech* 2012, 1819:97–103.

103. Mao X, Chen S, Li A, Zhai C, Jing R: Novel NAC Transcription Factor TaNAC67 confers enhanced multi-abiotic stress tolerances in *Arabidopsis*. *PLoS One* 2014, 9:e84359.

104. Chen X, Wang Y, Lv B, Li J, Luo L, Lu S, Zhang X, Ma H, Ming F: The NAC family transcription factor OsNAP confers abiotic stress response through the ABA pathway. *Plant Cell Physiol* 2014.

105. Meijer AH, Scarpella E, Van Dijk EL, Qin L, Taal AJC, Rueb S, Harrington SE, McCouch SR, Schilperoort RA, Hoge JHC: Transcriptional repression by Oshox1, a novel homeodomain leucine zipper protein from rice. *Plant J* 1997, 11:263–276.

106. Lee Y-H, Chun J-Y: A new homeodomain-leucine zipper gene from Arabidopsis thaliana induced by water stress and abscisic acid treatment. *Plant Mol Biol* 1998, 37:377–384.

107. Scarpella E, Rueb S, Boot KJ, Hoge JH, Meijer AH: A role for the rice homeobox gene Oshox1 in provascular cell fate commitment. *Development* 2000, 127:3655–3669.

108. Scarpella E, Boot KJM, Rueb S, Meijer AH: The procambium specification gene Oshox1 promotes polar auxin transport capacity and reduces its sensitivity toward inhibition. *Plant Physiol* 2002, 130:1349–1360.

109. Chen L, Zhou X, Li W, Chang W, Zhou R, Wang C, Sha A, Shan Z, Zhang C, Qiu D, Yang Z, Chen S: Genome-wide transcriptional analysis of two soybean genotypes under dehydration and rehydration conditions. *BMC Genomics* 2013, 14:687.

110. Di Cristina M, Sessa G, Dolan L, Linstead P, Baima S, Ruberti I, Morelli G: The Arabidopsis Athb-10 (GLABRA2) is an HD-Zip protein required for regulation of root hair development. *Plant J* 1996, 10:393–402.

111. Wang Y, Zhang W, Li K, Sun F, Han C, Li X: Salt-induced plasticity of root hair development is caused by ion disequilibrium in Arabidopsis thaliana. *J Plant Res* 2008, 121:87–96.

112. Wang Y, Li X: Salt stress-induced cell reprogramming, cell fate switch and adaptive plasticity during root hair development in Arabidopsis. *Plant Signal Behav* 2008, 3:436–438.

113. The Lens. [http://www.lens.org/lens/biological_search].

114. Jefferson OA, Kollhofer D, Ehrich TH, Jefferson RA: Transparency tools in gene patenting for informing policy and practice. *Nat Biotech* 2013, 31:1086–1093.

115. Palena CM, Gonzalez DH, Chan RL: A monomer-dimer equilibrium modulates the interaction of the sunflower homeodomain leucine-zipper protein Hahb-4 with DNA. *Biochem J* 1999, 341:81–87.

116. Palena CM, Tron AE, Bertoncini CW, Gonzalez DH, Chan RL: Positively charged residues at the N-terminal arm of the homeodomain are required for efficient DNA binding by homeodomain-leucine zipper proteins. *J Mol Biol* 2001, 308:39–47.

117. Johannesson H, Wang Y, Engstrom P: DNA-binding and dimerization preferences of Arabidopsis homeodomain-leucine zipper transcription factors in¬†vitro. *Plant Mol Biol* 2001, 45:63–73.

118. Arce AL, Raineri J, Capella M, Cabello JV, Chan RL: Uncharacterized conserved motifs outside the HD-Zip domain in HD-Zip subfamily I transcription factors; a potential source of functional diversity. *BMC Plant Biol* 2011, 11:42.

119. Zhang B, Chen W, Foley RC, Buttner M, Singh KB: Interactions between distinct types of DNA binding proteins enhance binding to ocs element promoter sequences. *Plant Cell Online* 1995, 7:2241–2252.

120. Chen W, Chao G, Singh KB: The promoter of a H2O2-inducible, Arabidopsis glutathione S-transferase gene contains closely linked OBF- and OBP1-binding sites. *Plant J* 1996, 10:955–966.

121. Kang H-G, Foley RC, Oñate-Sánchez L, Lin C, Singh KB: Target genes for OBP3, a Dof transcription factor, include novel basic helix-loop-helix domain proteins inducible by salicylic acid. *Plant J* 2003, 35:362–372.

122. Park DH, Lim PO, Kim JS, Cho DS, Hong SH, Nam HG: The Arabidopsis COG1 gene encodes a Dof domain transcription factor and negatively regulates phytochrome signaling. *Plant J* 2003, 34:161–171.

123. Yanagisawa S, Sheen J: Involvement of Maize Dof Zinc finger proteins in tissue-specific and light-regulated gene expression. *Plant Cell Online* 1998, 10:75–89.

124. Papi M, Sabatini S, Altamura MM, Hennig L, Schafer E, Costantino P, Vittorioso P: Inactivation of the phloem-specific Dof Zinc finger GeneDAG1 affects response to light and integrity of the testa of Arabidopsis seeds. *Plant Physiol* 2002, 128:411–417.

125. De Paolis A, Sabatini S, De Pascalis L, Costantino P, Capone I: A rolB regulatory factor belongs to a new class of single zinc finger plant proteins. *Plant J* 1996, 10:215–223.

126. Kisu Y, Ono T, Shimofurutani N, Suzuki M, Esaka M: Characterization and expression of a new class of zinc finger protein that binds to silencer region of ascorbate oxidase gene. *Plant Cell Physiol* 1998, 39:1054–1064.

127. Washio K: Identification of Dof proteins with implication in the gibberellin-regulated expression of a peptidase gene following the germination of rice grains. *Biochim Biophys Acta (BBA) Gene Struct Exp* 2001, 1520:54–62.

128. Mena M, Cejudo FJ, Isabel-Lamoneda I, Carbonero P: A role for the DOF transcription factor BPBF in the regulation of gibberellin-responsive genes in barley aleurone. *Plant Physiol* 2002, 130:111–119.

129. Papi M, Sabatini S, Bouchez D, Camilleri C, Costantino P, Vittorioso P: Identification and disruption of an Arabidopsis zinc finger gene controlling seed germination. *Genes Dev* 2000, 14:28–33.

130. Gualberti G, Papi M, Bellucci L, Ricci I, Bouchez D, Camilleri C, Costantino P, Vittorioso P: Mutations in the Dof Zinc finger Genes DAG2 and DAG1 influence with opposite effects the germination of Arabidopsis seeds. *Plant Cell Online* 2002, 14:1253–1263.

131. Desveaux D, Marechal A, Brisson N: Whirly transcription factors: defense gene regulation and beyond. *Trends Plant Sci* 2005, 10:95–102.

132. Winicov I, Bastola DR: Transgenic overexpression of the transcription FactorAlfin1 enhances expression of the endogenous MsPRP2Gene in Alfalfa and improves salinity tolerance of the plants. *Plant Physiol* 1999, 120:473–480.

133. Winicov I: Alfin1 transcription factor overexpression enhances plant root growth under normal and saline conditions and improves salt tolerance in alfalfa. *Planta* 2000, 210:416–422.

134. UniProtKB. [http://www.uniprot.org/uniprot/].

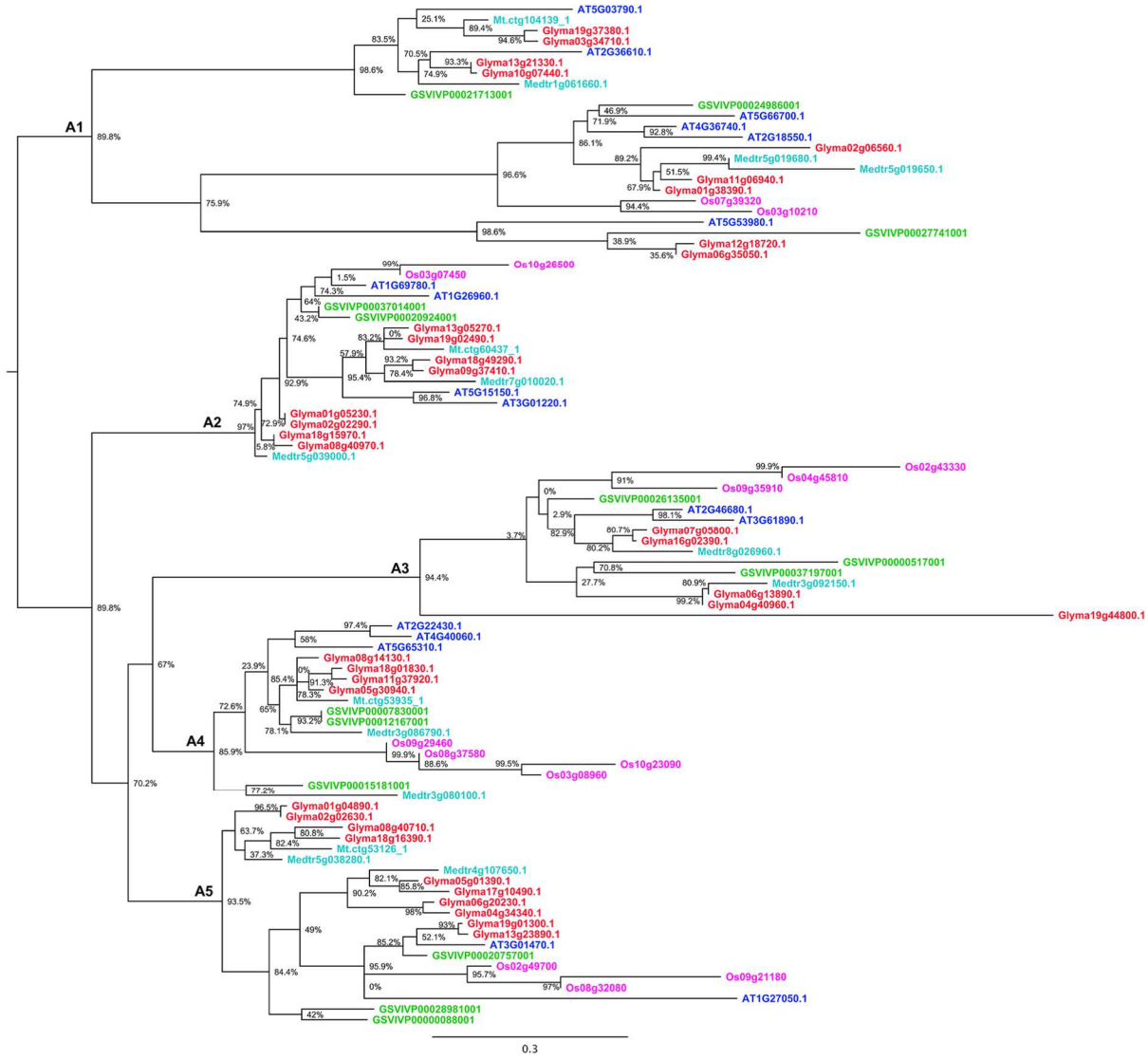

Figure 1. Phylogenetic relationships of HD-Zip I proteins from soybean, *Medicago*, *Arabidopsis*, grape and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1- A5 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies. Genes from each of the species are highlighted in different colors, soybean (red), *Medicago* (light blue), *Arabidopsis* (dark blue), grape (green), and rice (Pink).

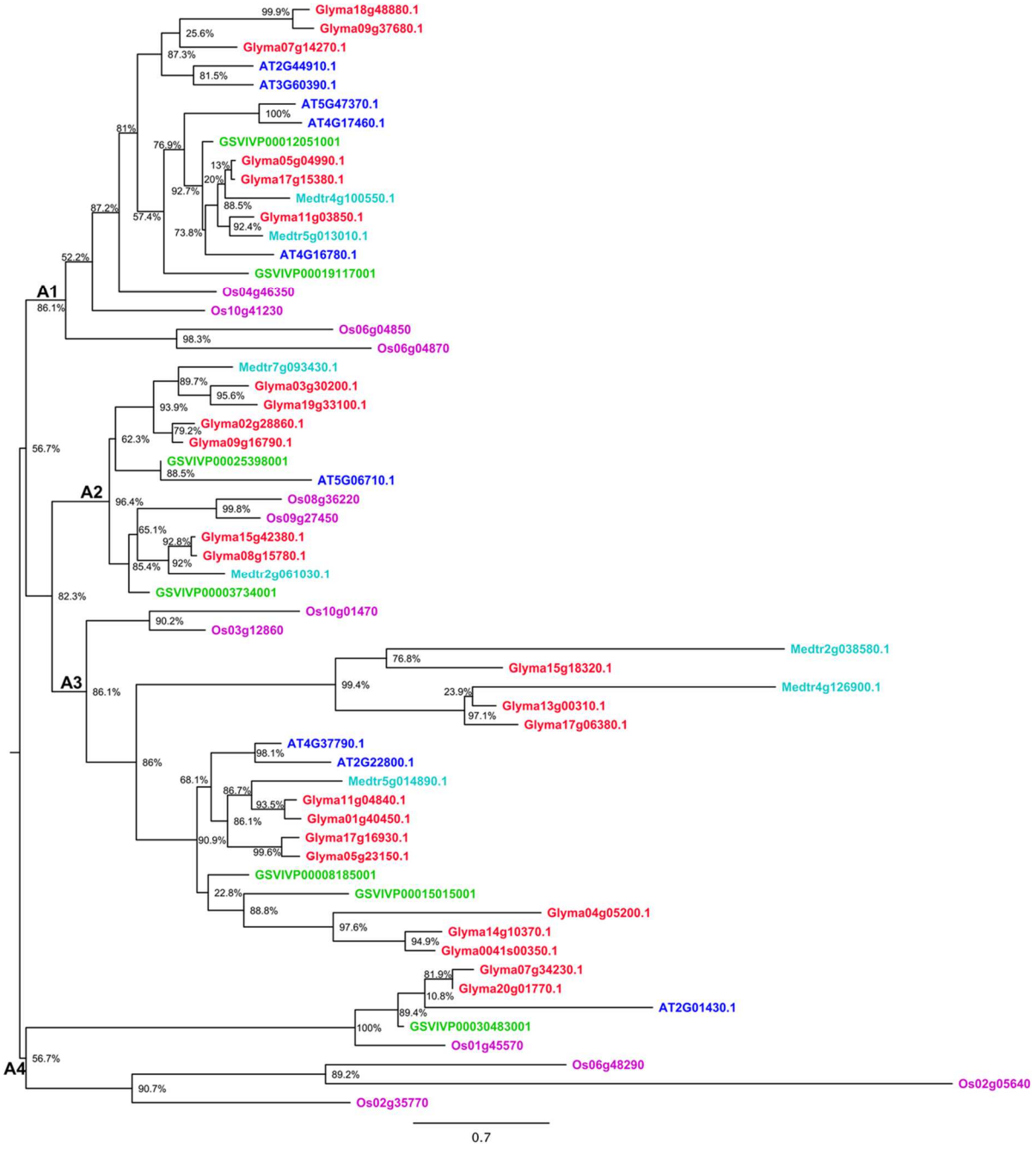

Figure 2. Phylogenetic relationships of HD-Zip II proteins from soybean, *Medicago*, *Arabidopsis*, grape and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1- A4 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies. Genes from each of the species are highlighted in different colors, soybean (red), *Medicago* (light blue), *Arabidopsis* (dark blue), grape (green), and rice (Pink).

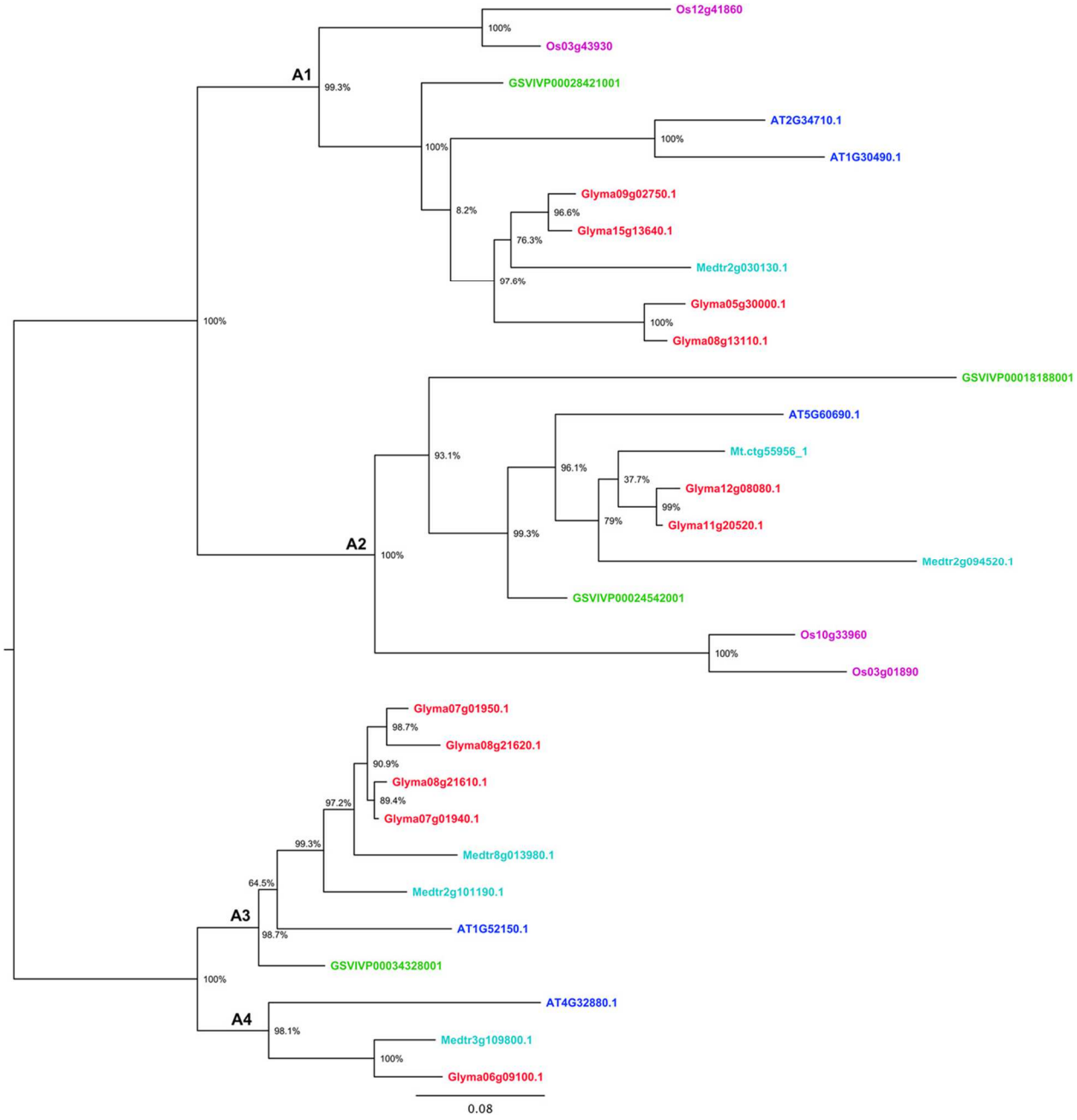

Figure 3. Phylogenetic relationships of HD-Zip III proteins from soybean, *Medicago*, *Arabidopsis*, grape and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1- A4 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies. Genes from each of the species are highlighted in different colors, soybean (red), *Medicago* (light blue), *Arabidopsis* (dark blue), grape (green), and rice (Pink).

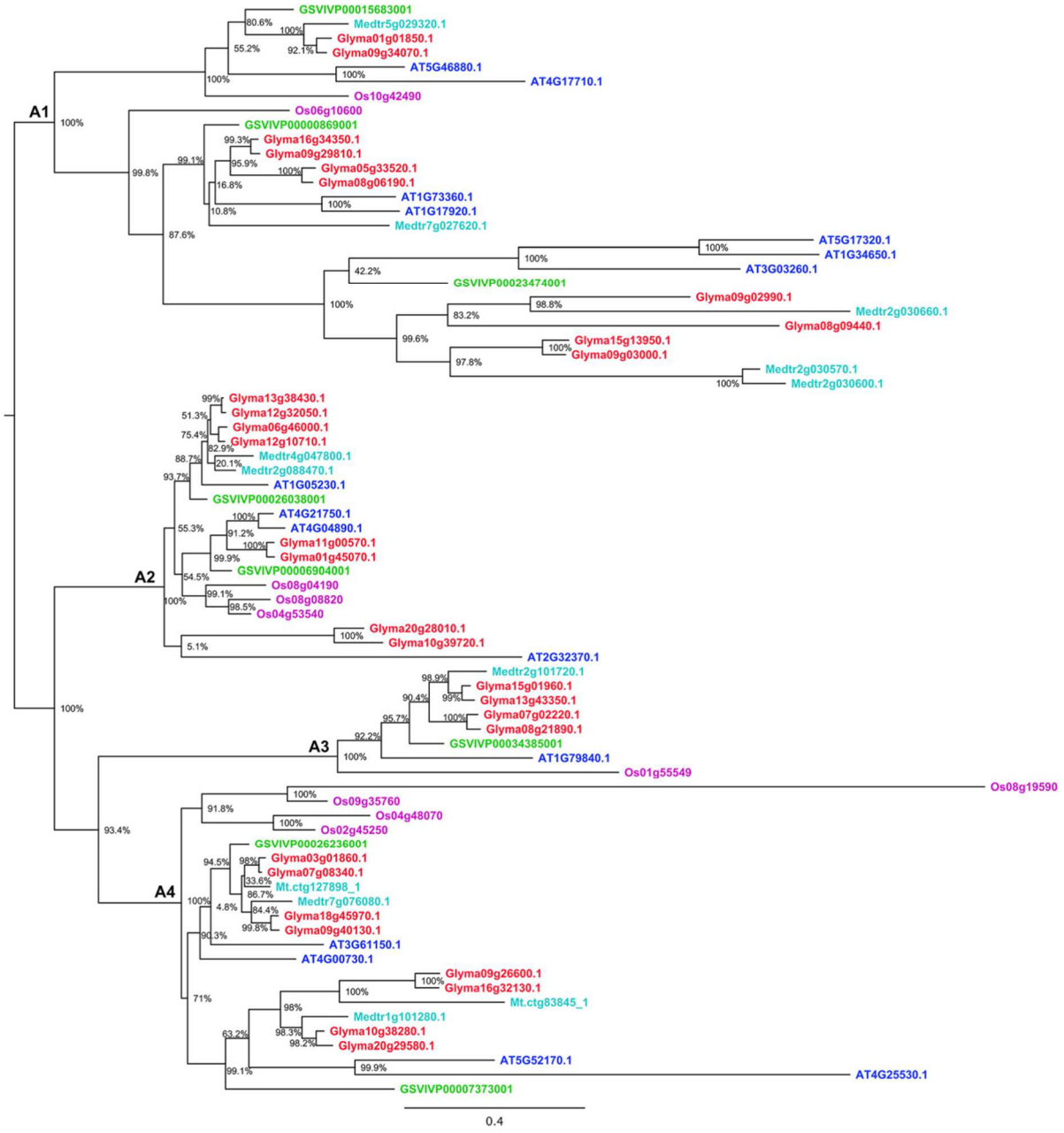

Figure 4. Phylogenetic relationships of HD-Zip IV proteins from soybean, *Medicago*, *Arabidopsis*, grape and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1- A4 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies. Genes from each of the species are highlighted in different colors, soybean (red), *Medicago* (light blue), *Arabidopsis* (dark blue), grape (green), and rice (Pink). Genes Medtr5g005600.1 and Os01g57890 belong to the angiosperm clade "A2." These two genes are not shown in the phylogeny because adding them significantly affects the topology.

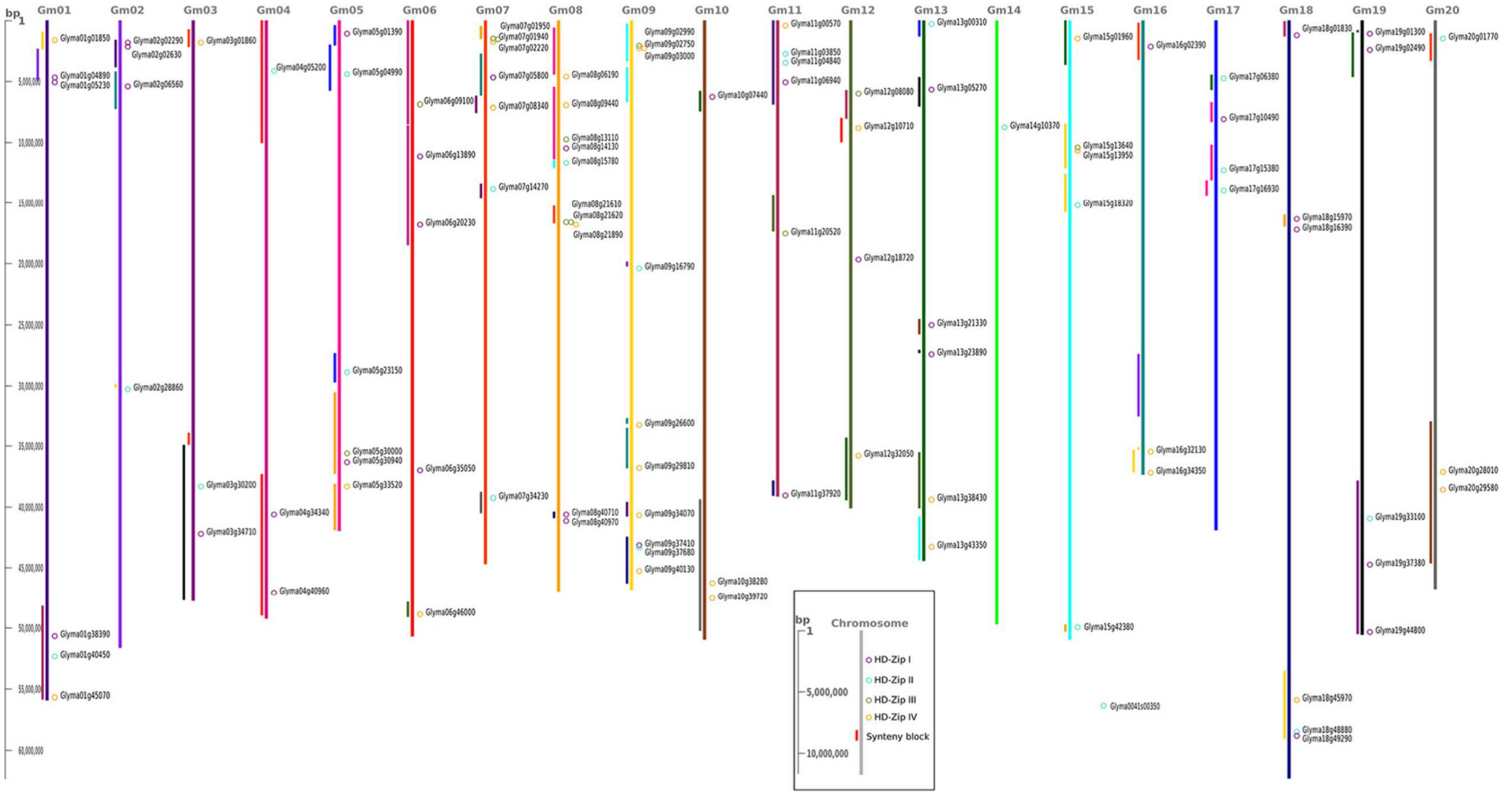

Figure 5. Chromosomal locations and synteny relationships of soybean HD-Zip genes. The chromosomal locations of the soybean HD-Zip genes were obtained from the GFF file of *Glycine max* assembly v1.01, annotation 1.09, and were displayed using chromosome visualization tool (CViT). All chromosomes and gene locations are shown to scale. Glyma0041s00350 located on scaffold 41 (149758–152298 bp) is included independently in the figure. The homoeologous gene pairs are identified with colored solid lines on the left side of the chromosomes. The chromosomes and the solid lines with identical colors are syntenic regions containing homoeologous genes. A detailed list of homoeologous HD-Zip genes is also provided in Additional file 13: Table S1. The HD-Zip I genes are indicated in yellow, HD-Zip II in purple, HD-Zip III in blue, and HD-Zip IV in green.

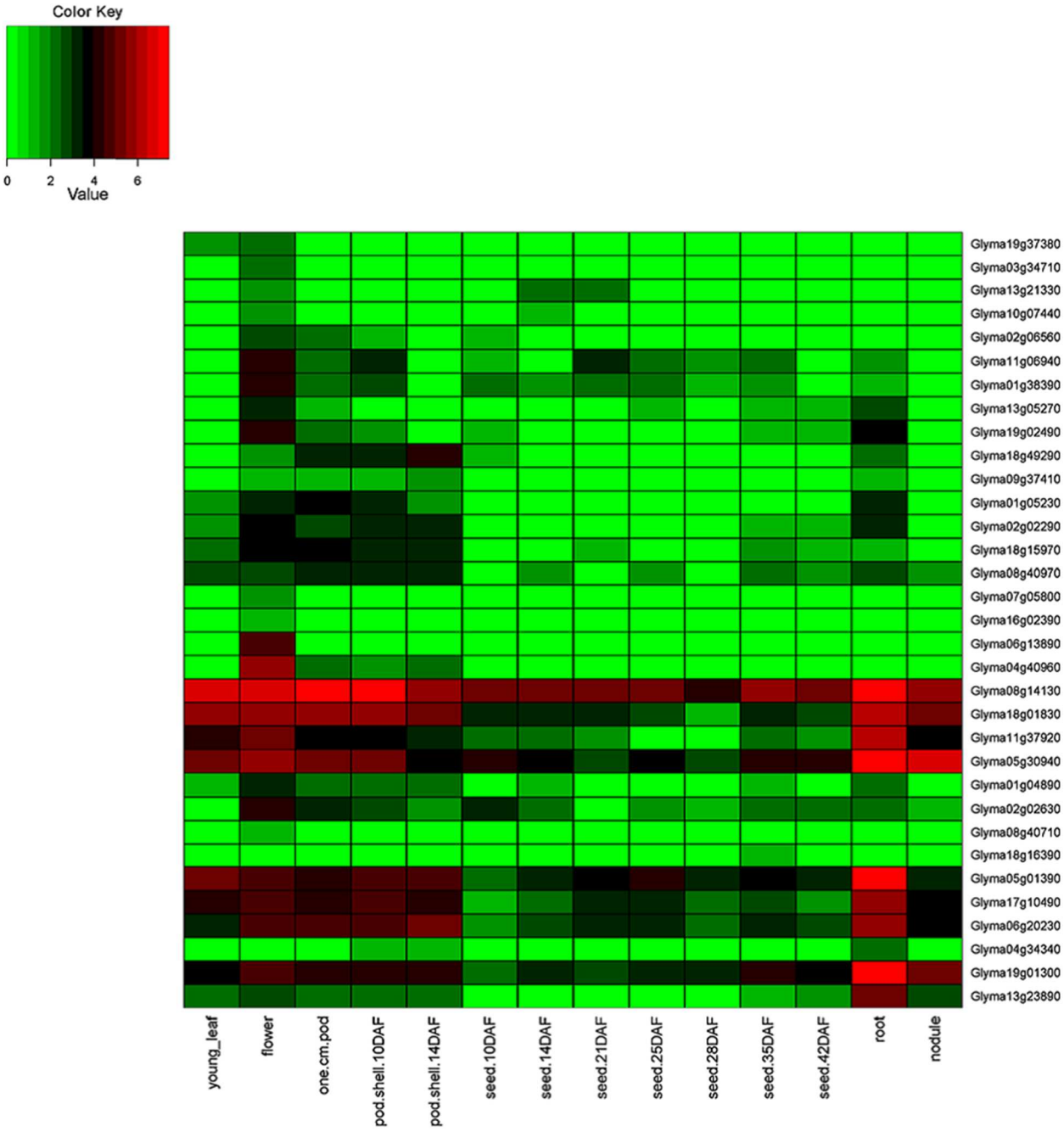

Figure 6. Expression profiles of HD-Zip I genes in 14 tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 1. The abbreviation "DAF" in the tissue label represents "Days after flowering."

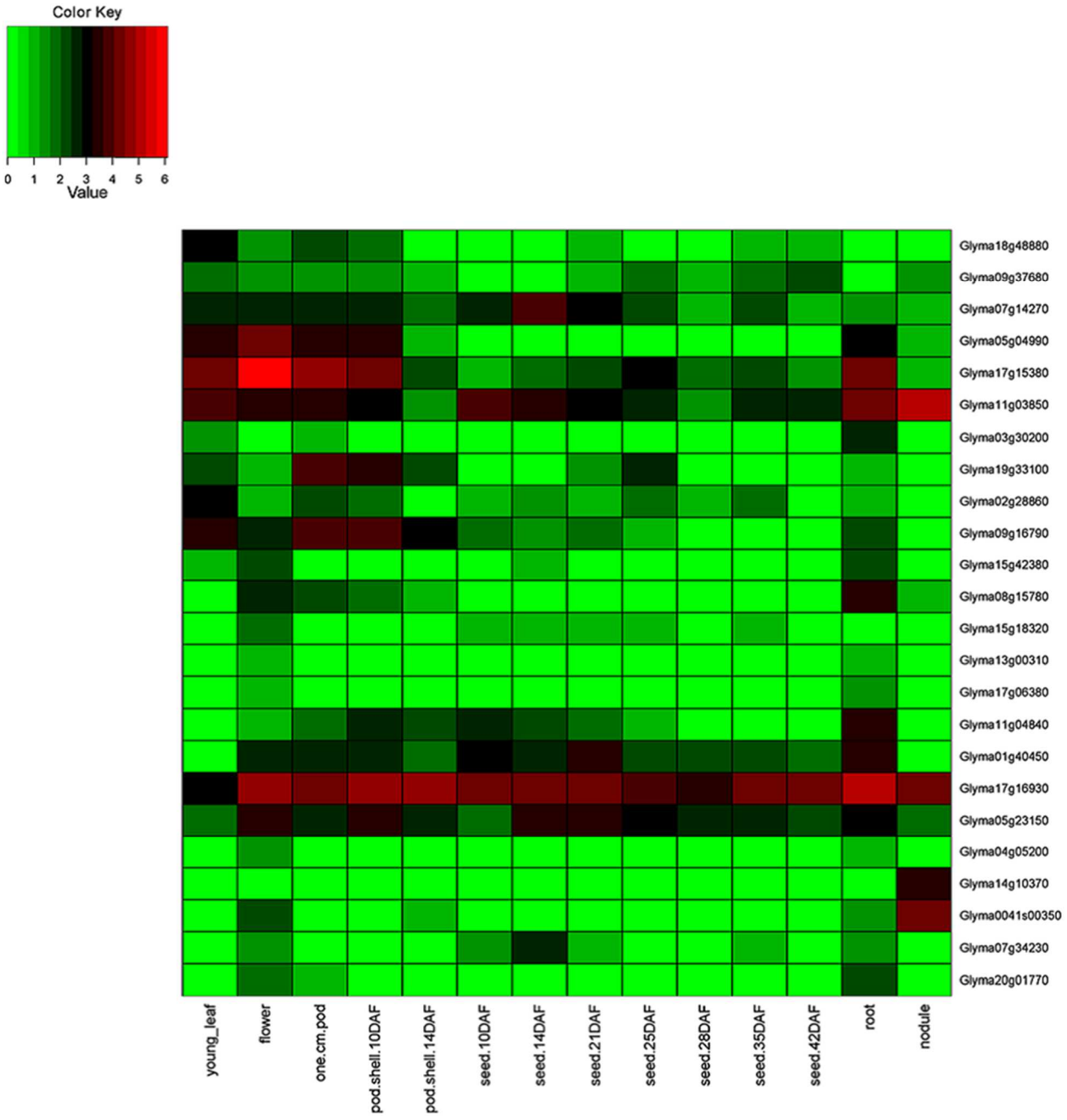

Figure 7. Expression profiles of HD-Zip II genes in 14 tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 2. The abbreviation "DAF" in the tissue label represents "Days after flowering."

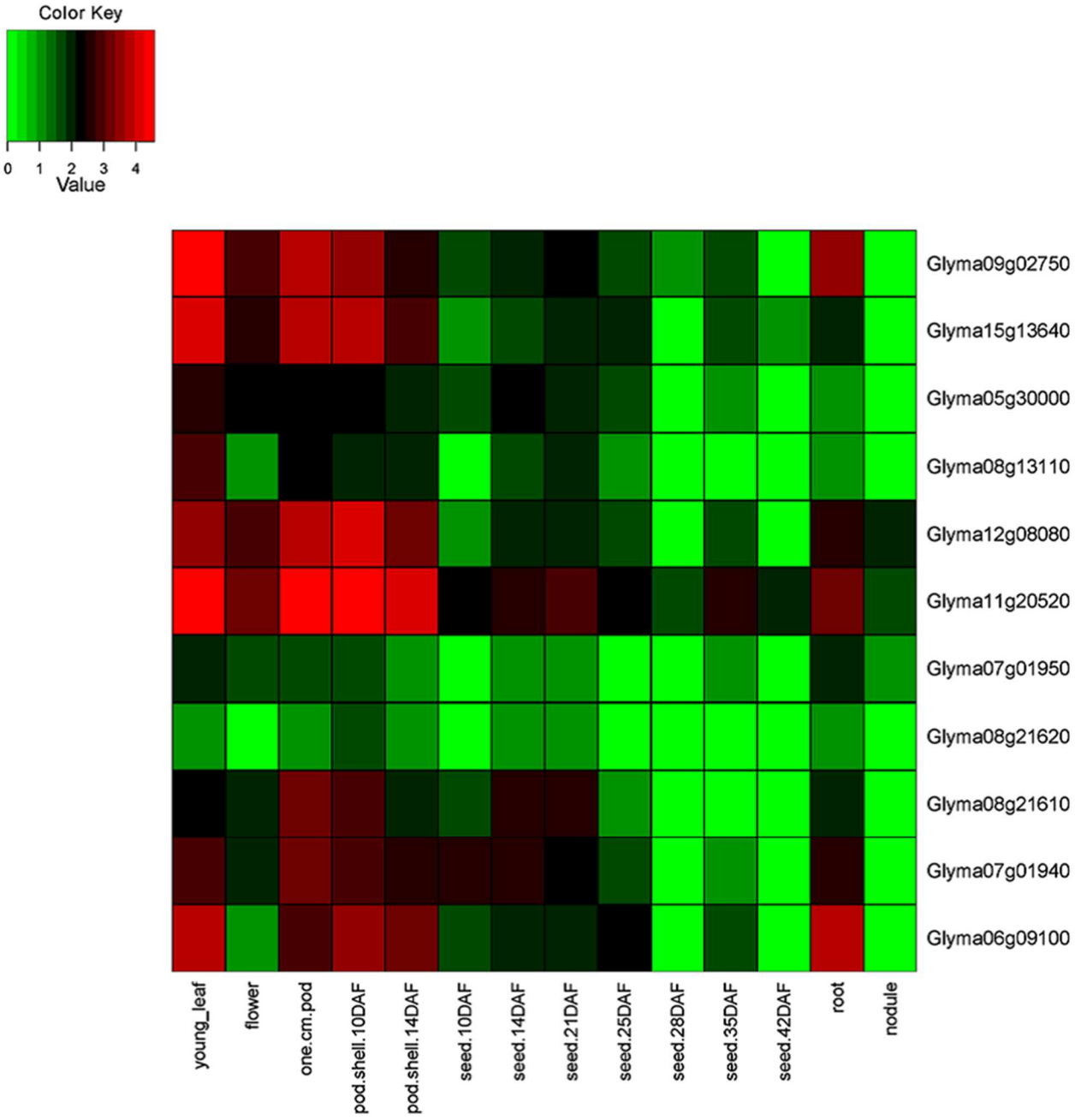

Figure 8. Expression profiles of HD-Zip III genes in 14 tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 3. The abbreviation "DAF" in the tissue label represents "Days after flowering."

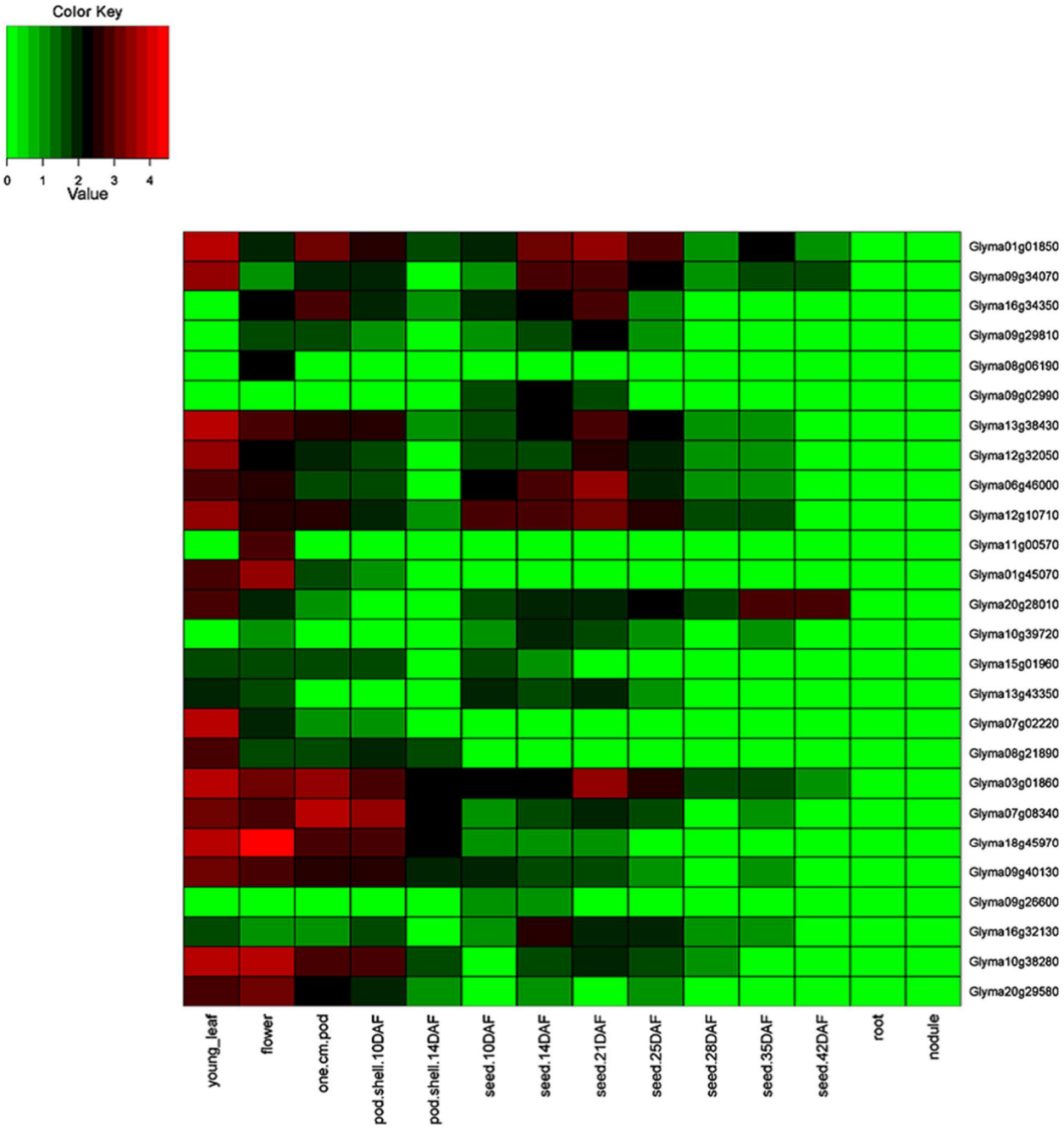

Figure 9. Expression profiles of HD-Zip IV genes in 14 tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 4. The abbreviation "DAF" in the tissue label represents "Days after flowering."

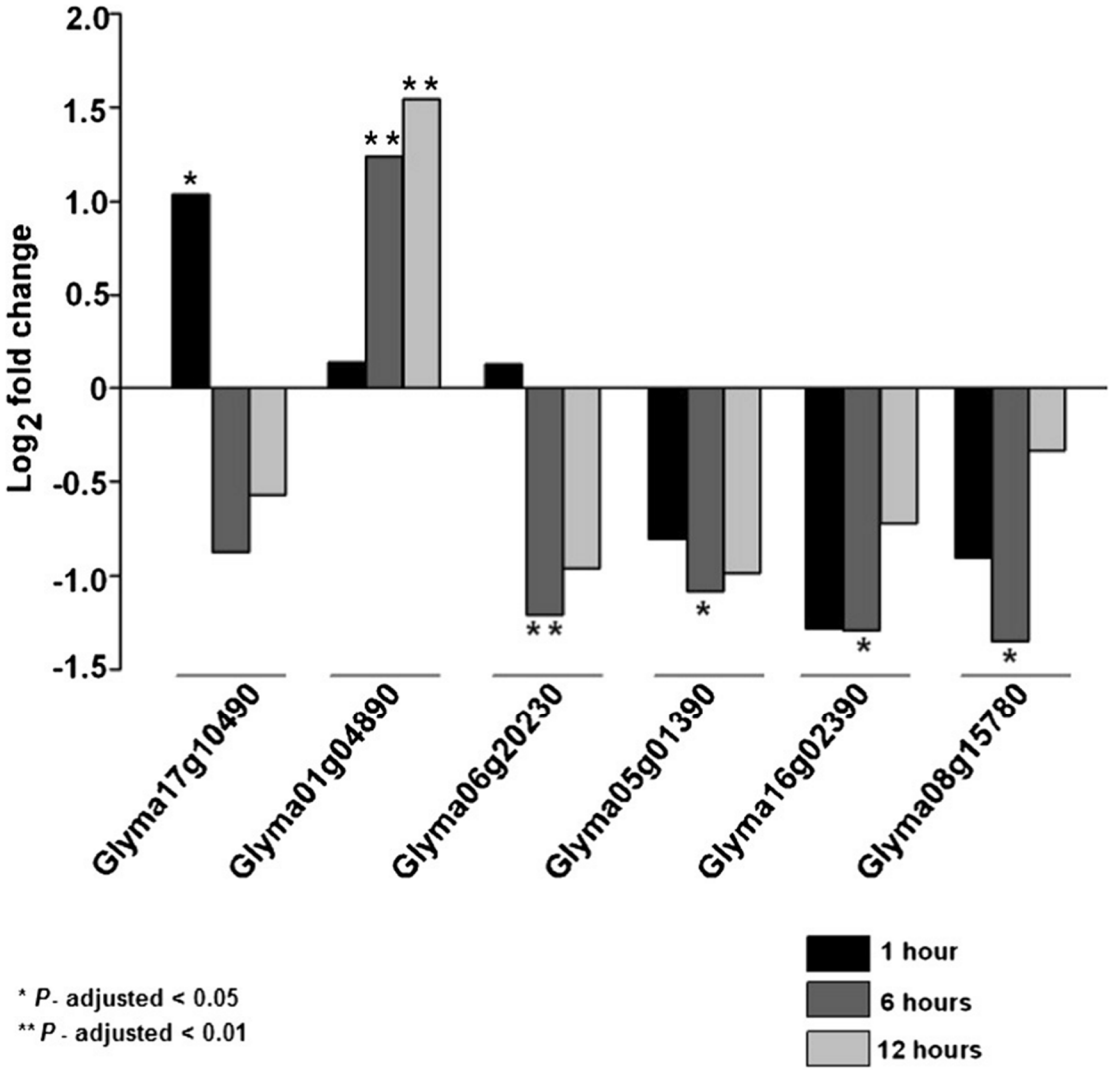

Figure 10. RNA-Seq based expression profiles of HD-Zip genes that are differentially expressed in at least one time point under dehydration stress. The HD-Zip genes responsive to dehydration stress at the first trifoliolate stage in the roots of soybean cv. Williams 82 at least at one time point (1, 6 or 12 hr) are shown. The criteria for differential expression includes, (1) *P*-value corrected for multiple testing correction using Benjamini and Hochberg [62] to be less than 0.05, (2) two folder or greater fold change, (3) residual variance quotients of both control and treatment samples be less than 20. The criterion (3) filters genes that have significant variation between replicates.

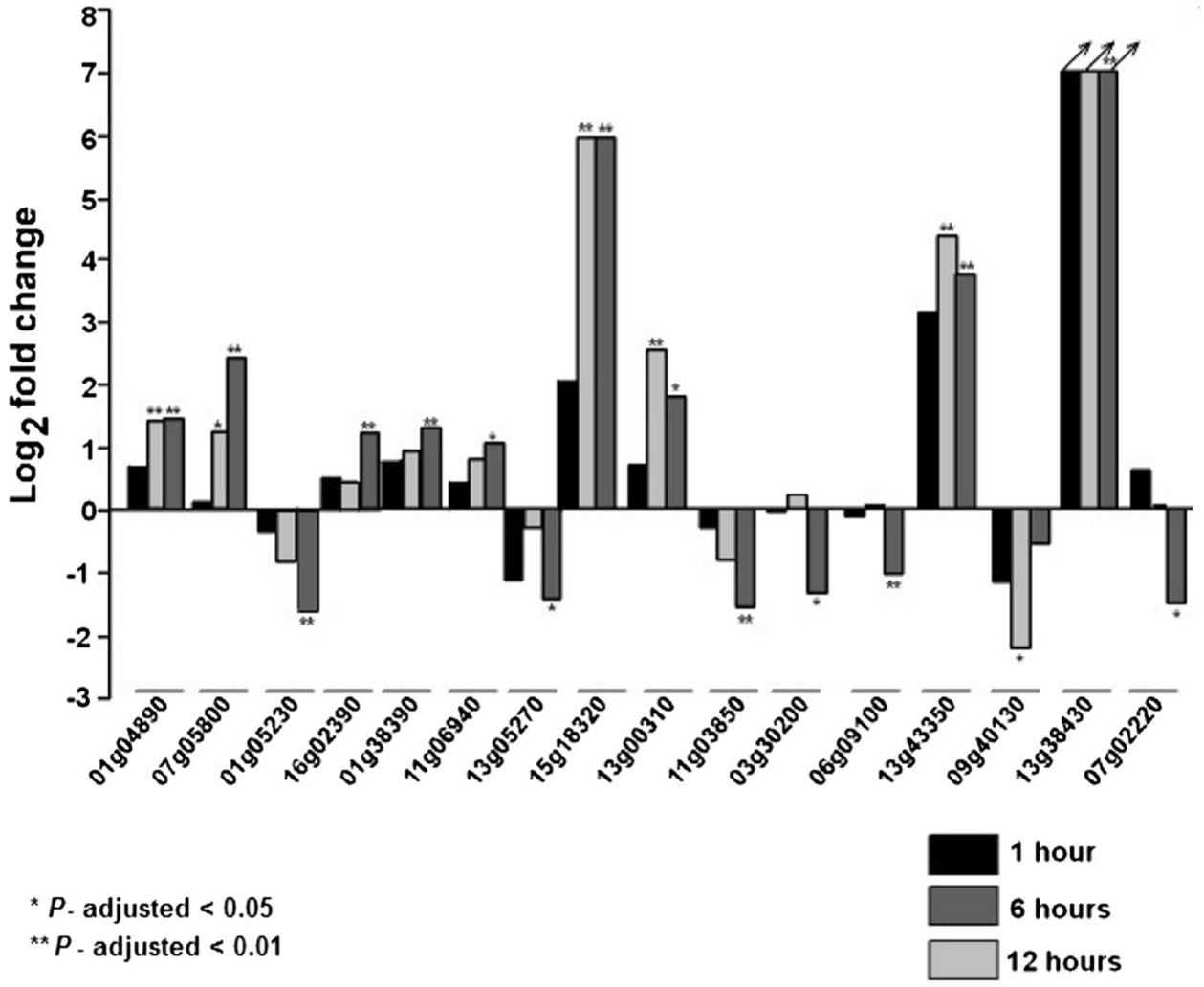

Figure 11. RNA-Seq based expression profiles of HD-Zip genes that are differentially expressed in at least one time point under salt stress. The HD-Zip genes responsive to salt stress at the first trifoliolate stage in the roots of soybean cv. Williams 82 at least at one time point (1, 6 or 12 hr) are shown. The criteria for differential expression includes, (1) *P*-value corrected for multiple testing correction using Benjamini and Hochberg [62] to be less than 0.05, (2) two folder or greater fold change, (3) residual variance quotients of both control and treatment samples be less than 20. The criterion (3) filters genes that have significant variation between replicates.

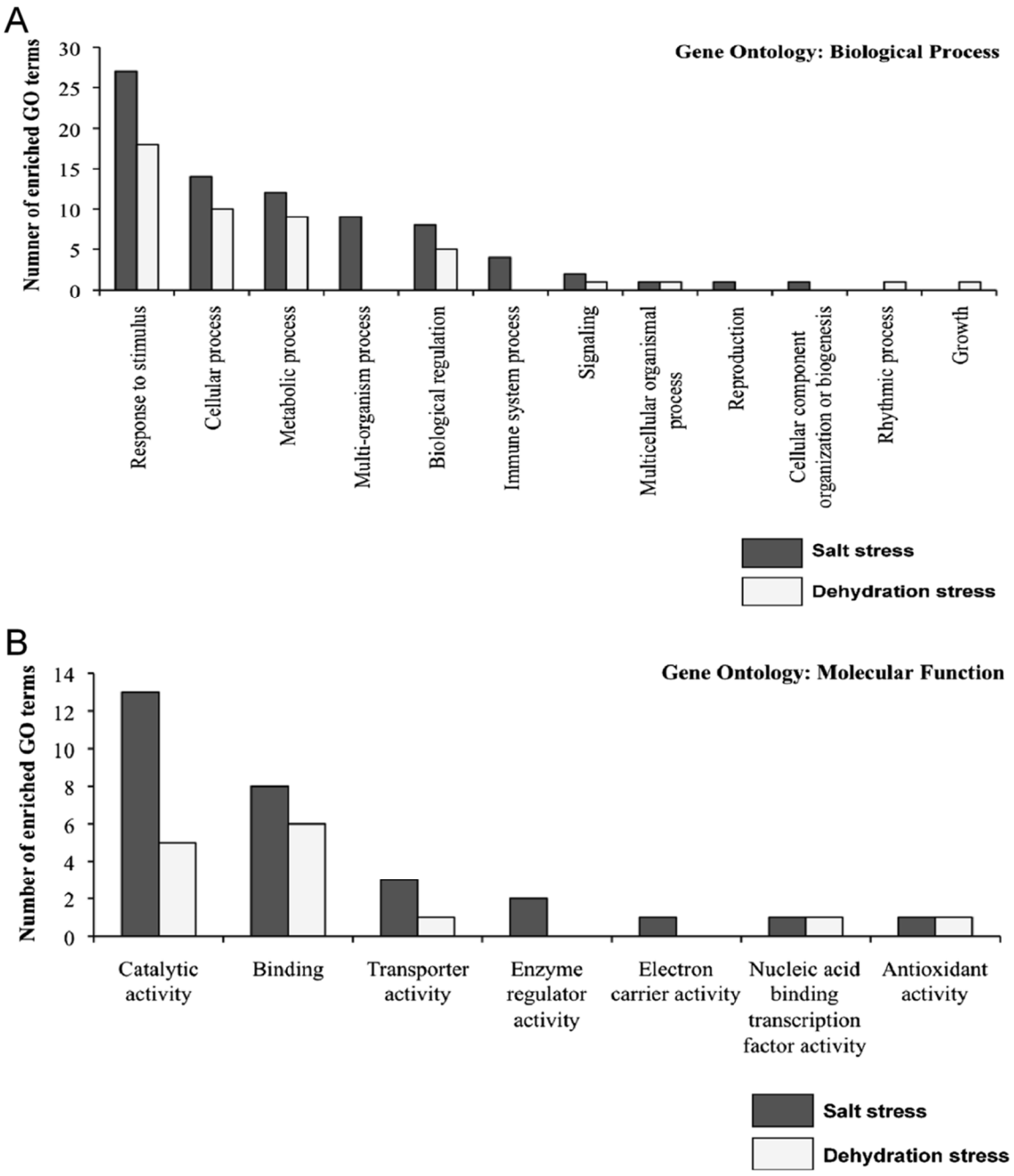

Figure 12. Gene ontology biological process (A) and molecular function (B) categories significantly (corrected P <0.05) overrepresented among differentially expressed genes under dehydration and salt stress. Differentially expressed genes under dehydration and salt stress were annotated using the top Arabidopsis hit, and then screened for overrepresented GO terms against all soybean genes using Fisher's exact test [66] and Bonferroni [67] corrected significance value of less than 0.05 (Additional file 31: Table S11). The overrepresented GO terms were enriched at the second level using BLAST2GO v.2.7.1 [68] and are shown in the figure.

Table 1. Number of HD-Zip genes observed (O), expected (E) and retained (R) among five angiosperm species

	HD-Zip I (5)[a]			HD-Zip II (4)[a]			HD-Zip III (4)[a]			HD-Zip IV (4)[a]			Total - Each species		
Species	**O**	**E**	**R (%)**	**O**	**E**	**R (%)**	**O**	**E**	**R (%)**	**O**	**E**	**R (%)**	**O**	**E**	**R (%)**
Arabidopsis thaliana (12)[b]	17	60	28.3	9	48	18.8	5	48	10.4	16	48	33.3	47	204	23.0
Vitis vinifera (3)[b]	14	15	93.3	7	12	58.3	4	12	33.3	8	12	66.7	33	51	64.7
Glycine max (12)[b]	36	60	60.0	24	48	50.0	11	48	22.9	30	48	62.5	101	204	49.5
Medicago truncatula (6)[b]	15	30	50.0	7	24	29.2	6	24	25.0	13	24	54.2	41	102	40.2
Oryza sativa (4)[b]	14	20	70.0	12	16	75.0	4	16	25.0	11	16	68.8	41	68	60.3
Total - Among five species	96	185	51.9	59	148	39.9	30	148	20.3	78	148	52.7	263	629	41.8

[a]Number of ancient angiosperm clades observed in each HD-Zip subfamily.
[b]Number of genes expected in each ancient angiosperm clade based on the history of whole genome duplication events.

Table 2 Experimental set-up and summary of read-count data from RNA-Seq analysis

Treatment	Time point (hr)	Replicate (#)	Lane on HiSeq 2000	Total reads	Uniquely mapped reads	Uniquely mapped reads (%)
Control	0	1	4	10,150,369	8,047,650	79.3%
Control	0	2	4	11,849,953	9,207,421	77.7%
Control	0	3	3	11,272,789	9,164,808	81.3%
Dehydration	1	1	3	6,875,669	5,571,406	81.0%
Dehydration	1	2	3	6,744,882	4,948,665	73.4%
Dehydration	1	3	3	8,650,675	6,612,402	76.4%
Dehydration	6	1	3	11,828,271	9,624,573	81.4%
Dehydration	6	2	3	11,355,361	8,599,941	75.7%
Dehydration	6	3	4	10,038,099	8,009,975	79.8%
Dehydration	12	1	4	9,270,260	7,194,931	77.6%
Dehydration	12	2	4	5,555,797	4,050,429	72.9%
Dehydration	12	3	4	5,105,827	4,087,108	80.0%
Salt	1	1	2	38,214,261	30,683,738	80.3%
Salt	1	2	2	9,046,880	7,428,387	82.1%
Salt	1	3	2	9,423,474	7,202,416	76.4%
Salt	6	1	2	7,445,356	5,580,211	74.9%
Salt	6	2	2	5,890,968	4,422,058	75.1%
Salt	6	3	2	25,296,306	14,687,424	58.1%
Salt	12	1	2	7,481,184	5,253,438	70.2%
Salt	12	2	3	14,201,579	11,170,436	78.7%
Salt	12	3	4	13,124,265	9,645,045	73.5%
			Total	238,822,225	181,192,462	75.9%
			Average	11,372,487	8,628,212	76.5%

Table 3. Transcription factor class significantly (corrected P <0.05) overrepresented among the differentially expressed genes under dehydration and salt stress

Transcription factor class	Genome count	Salt stress expression count	Corrected *P*-value	Dehydration stress expression count	Corrected *P*-value	Role in abiotic stress response
WRKY	197	82	2.52E-21	34	3.45E-03	[90-92]
AP2-EREBP	381	111	1.67E-14	75	2.88E-10	[93-95]
ZIM	24	20	9.55E-13	16	1.28E-10	[96-98]
C2C2 (Zn) CO-like	72	33	9.94E-10	26	3.77E-09	[99-101]
NAC	208	49	3.48E-03	NA	NA	[102-104]

Table 4. Plant transcription factor binding sites significantly (P <0.05) overrepresented in the promoters of HD-Zip genes belonging to each of the subfamilies

	Motif #	[1]TFBS	[2]Count	[3]Proportion	[4]TFBS_Dehydration	[5]TFBS_Salt	[6]TF_Class	[7]DE_Dehydration	[8]DE_Salt
HD-Zip I	M00354	Dof3	33	91.7	-	-	Dof	+	+
	M00700	ROM	31	86.1	+	-	bZIP	+	+
	M01136	Dof	29	80.6	-	-	Dof	+	+
	M00353	Dof2	27	75.0	-	-	Dof	+	+
[9]HD-Zip II	M00479	Alfin1	21	91.3	-	-	PHD	+	+
	M01136	Dof	20	87.0	-	-	Dof	+	+
	M00354	Dof3	19	82.6	-	-	Dof	+	+
	M00440	CG1	18	78.3	-	-	CAMTA	+	+
	M00506	LIM1	18	78.3	-	-	LIM	-	+
	M00502	TEIL	17	73.9	-	+	[10]AP2-EREBP	+	+
	M00653	OCSBF-1	17	73.9	+	+	bZIP	+	+
	M00788	EmBP-1b	17	73.9	+	+	bZIP	+	+
	M01128	SED	17	73.9	-	-	DOF	+	+
	M00942	CPRF-1	16	69.6	+	-	bZIP	+	+
	M00948	PCF2	16	69.6	+	-	TCP	+	+
	M00443	Opaque-2	14	60.9	+	+	bZIP	+	+
	M01133	AG	14	60.9	-	-	MADS	+	+
	M00660	RITA-1	13	56.5	+	+	bZIP	+	+
	M01130	PBF	13	56.5	-	-	Dof	+	+
	M01054	bHLH66	12	52.2	+	+	bHLH	+	+
	M00503	ATHB-5	11	47.8	+	-	HD-Zip I	+	+
	M00434	PIF3	10	43.5	+	+	bHLH	+	+
HD-Zip III	M00479	Alfin1	10	90.9	-	-	PHD	+	+
	M00438	ARF	9	81.8	-	-	ARF	+	+
	M01021	ID1	9	81.8	+	-	C2H2 - zinc	+	+
	M01126	BPC1	8	72.7	-	-	BBR/BPC	-	-
	M00948	PCF2	7	63.6	+	-	TCP	+	+
	[11] M00151	AG	7	63.6	-	-	MADS	+	+
	M00820	HAHB-4	6	54.5	-	-	HD-Zip I	+	+
	[12] M01061	AGL2	6	54.5	-	-	MADS	+	+
	M00392	AGL3	5	45.5	-	-	MADS	+	+
	M00949	AGL15	5	45.5	-	-	MADS	+	+
HD-Zip IV	[13] M00355	PBF	29	96.7	+	-	Dof	+	+
	M00438	ARF	25	83.3	-	-	ARF	+	+
	M01126	BPC1	25	83.3	-	-	BBR/BPC	-	-
	M01136	Dof	25	83.3	-	-	Dof	+	+
	M01128	SED	23	76.7	-	-	DOF	+	+
	M01021	ID1	22	73.3	+	-	C2H2 - zinc	+	+
	M00702	SPF1	20	66.7	-	+	[10]WRKY	+	+
	M00654	OSBZ8	15	50.0	+	+	bZIP	+	+
	M00089	Athb-1	11	36.7	+	-	HD-Zip I	+	+

[1]TFBS: Transcription factor binding site (TFBS) significantly (P <0.05, motif score >5) overrepresented in the promoters of HD-Zip genes.
[2]Count: Number of HD-Zip genes within a subfamily that contain the TFBS significantly overrepresented in their promoters.
[3]Proportion: Percentage of HD-Zip genes within a subfamily that contain the TFBS overrepresented in their promoters.
[4]TFBS_Dehydration: " + " indicates that the respective TFBS is overrepresented in the promoters of genes that were differentially expressed under dehydration stress, and "-" represents not overrepresented.
[5]TFBS_Salt: " + " indicates that the respective TFBS is overrepresented in the promoters of genes that were differentially expressed under salt stress, and "-" represents not overrepresented.
[6]TF_Class: The membership of TFBS to a particular transcription factor (TF) class based on TRANSFAC [72] and UniprotKB [134].
[7]DE_Dehydration: " + " indicates members of the respective TF class are differentially expressed (DE) under dehydration stress, and "-" indicates otherwise.
[8]DE_Salt: " + " indicates members of the respective TF class are DE under salt stress, and "-" indicates otherwise.
[9]Although the HD-Zip II subfamily has 24 genes, the proportion is calculated using 23 genes. HD-Zip II gene Glyma05g23150 was excluded from the promoter analysis due to the selection criteria utilized (see methods for promoter selection criteria).
[10] TF class significantly (corrected P <0.05) overrepresented in the DE genes under dehydration and salt stress.
[11]AG TFBS has multiple motif identifiers - M00151, M01063, M01133, and M00950. Counts of AG TFBS's irrespective of the identifier# were summed to estimate total count and proportion.
[12]AGL2 TFBS has two motif identifiers - M01061 and M01062. Counts of AGL2 TFBS's irrespective of the identifier# were summed to estimate total count and proportion.
[13]PBF TFBS has two motif identifiers - M00355 and M01130. Counts of PBF TFBS's irrespective of the identifier# were summed to estimate total count and proportion.

Additional Files

Additional file 1: Figure S1. Sequence logo of HD-Zip I displaying the conserved residues in HMM alignment.

Additional file 2: Figure S2. Sequence logo of HD-Zip II displaying the conserved residues in HMM alignment.

Additional file 3: Figure S3. Sequence logo of HD-Zip III displaying the conserved residues in HMM alignment.

Additional file 4: Figure S4. Sequence logo of HD-Zip IV displaying the conserved residues in HMM alignment.

Additional file 5: Figure S5. Phylogenetic relationships of HD-Zip I proteins from soybean, *Medicago*, *Arabidopsis*, grape, poplar, cucumber, maize and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1- A5 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The letters are ordered for consistency with the phylogeny in Figure 1. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies.

Additional file 6: Figure S6. Phylogenetic relationships of HD-Zip II proteins from soybean, *Medicago*, *Arabidopsis*, grape, poplar, maize and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1- A4 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The letters are ordered for consistency with the phylogeny in Figure 2. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies.

Additional file 7: Figure S7. Phylogenetic relationships of HD-Zip III proteins from soybean, *Medicago*, *Arabidopsis*, grape, poplar, maize and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1- A4 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The letters are ordered for consistency with the phylogeny in Figure 3. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies.

Additional file 8: Figure S8. Phylogenetic relationships of HD-Zip IV proteins from soybean, *Medicago*, *Arabidopsis*, grape, poplar, cucumber, maize and rice. The phylogenetic tree was built using the maximum likelihood method implemented in PhyML. The letters A1-

A4 represent ancient angiosperm clades, based on whole genome duplication events, and the copy number of genes from each of the species. The letters are ordered for consistency with the phylogeny in Figure 4. The branch support values estimated using approximate likelihood ratio test (aLRT) are displayed in percentages. Rooting of the tree was inferred from Ariel et al. [1], angiosperm clade composition, and outgroup sequences from other subfamilies. Genes Medtr5g005600.1 and Os01g57890 belong to the angiosperm clade "A2." These two genes are not shown in the phylogeny because adding them significantly affects the topology.

Additional file 9: Figure S9. Gene structure of HD-Zip I genes showing the exon-intron structure.

Additional file 10: Figure S10. Gene structure of HD-Zip II genes showing the exon-intron structure.

Additional file 11: Figure S11. Gene structure of HD-Zip III genes showing the exon-intron structure.

Additional file 12: Figure S12. Gene structure of HD-Zip IV genes showing the exon-intron structure.

Additional file 13: Table S1. List of homoeologous soybean HD-Zip genes.

Additional file 14: Figure S13. Expression profiles of HD-Zip I genes in seven tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 1. The abbreviation "SAM" in the tissue label represents "shoot apical meristem."

Additional file 15: Figure S14. Expression profiles of HD-Zip II genes in seven tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 2. The abbreviation "SAM" in the tissue label represents "shoot apical meristem."

Additional file 16: Figure S15. Expression profiles of HD-Zip III genes in seven tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 3. The abbreviation "SAM" in the tissue label represents "shoot apical meristem."

Additional file 17: Figure S16. Expression profiles of HD-Zip IV genes in seven tissues of soybean. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 4. The abbreviation "SAM" in the tissue label represents "shoot apical meristem."

Additional file 18: Figure S17. Expression profiles of HD-Zip I genes in mock-inoculated and *Bradyrhizobium japonicum*-infected root hair cells harvested at 12, 24, and 48 hr after inoculation (HAI), and stripped roots harvested at 48 HAI with *B. japonicum*. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 1. The abbreviation RH_UN and RH_IN in the tissue label represent mock-inoculated and *B. japonicum* infected root hair cells respectively. The sample RS_48HAI_IN represents stripped roots harvested at 48 HAI with *B. japonicum*.

Additional file 19: Figure S18. Expression profiles of HD-Zip II genes in mock-inoculated and *Bradyrhizobium japonicum*-infected root hair cells harvested at 12, 24, and 48 hr after inoculation (HAI), and stripped roots harvested at 48 HAI with *B. japonicum*. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 2. The abbreviation RH_UN and RH_IN in the tissue label represent mock-inoculated and *B. japonicum* infected root hair cells respectively. The sample RS_48HAI_IN represents stripped roots harvested at 48 HAI with *B. japonicum*.

Additional file 20: Figure S19. Expression profiles of HD-Zip III genes in mock-inoculated and *Bradyrhizobium japonicum*-infected root hair cells harvested at 12, 24, and 48 hr after inoculation (HAI), and stripped roots harvested at 48 HAI with *B. japonicum*. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 3. The abbreviation RH_UN and RH_IN in the tissue label represent mock-inoculated and *B. japonicum* infected root hair cells respectively. The sample RS_48HAI_IN represents stripped roots harvested at 48 HAI with *B. japonicum*.

Additional file 21: Figure S20. Expression profiles of HD-Zip IV genes in mock-inoculated and *Bradyrhizobium japonicum*-infected root hair cells harvested at 12, 24, and 48 hr after inoculation (HAI), and stripped roots harvested at 48 HAI with *B. japonicum*. The Reads/Kb/Million (RPKM) normalized values of expressed genes was $\log_2$-transformed and visualized as heatmaps. Genes in the heatmap are ordered for consistency with the phylogeny in Figure 4. The abbreviation RH_UN and RH_IN in the tissue label represent mock-inoculated and *B. japonicum* infected root hair cells respectively. The sample RS_48HAI_IN represents stripped roots harvested at 48 HAI with *B. japonicum*.

Additional file 22: Table S2. Soybean genes differentially expressed under dehydration stress at 1 hr. The table includes mean expression values under control and stress conditions; fold change and $\log_2$ fold change values, *P*-values and adjusted *P*-values, and residual variance quotients of control and treatment samples. See Material and Methods for the criteria of differential expression.

Additional file 23: Table S3. Soybean genes differentially expressed under dehydration stress at 6 hr. The table includes mean expression values under control and stress conditions; fold change and $\log_2$ fold change values, *P*-values and adjusted *P*-values, and residual

variance quotients of control and treatment samples. See Material and Methods for the criteria of differential expression.

Additional file 24: Table S4. Soybean genes differentially expressed under dehydration stress at 12 hr. The table includes mean expression values under control and stress conditions; fold change and $\log_2$ fold change values, *P*-values and adjusted *P*-values, and residual variance quotients of control and treatment samples. See Material and Methods for the criteria of differential expression.

Additional file 25: Table S5. Soybean genes differentially expressed under salt stress at 1 hr. The table includes mean expression values under control and stress conditions; fold change and $\log_2$ fold change values, *P*-values and adjusted *P*-values, and residual variance quotients of control and treatment samples. See Material and Methods for the criteria of differential expression.

Additional file 26: Table S6. Soybean genes differentially expressed under salt stress at 6 hr. The table includes mean expression values under control and stress conditions; fold change and $\log_2$ fold change values, *P*-values and adjusted *P*-values, and residual variance quotients of control and treatment samples. See Material and Methods for the criteria of differential expression.

Additional file 27: Table S7. Soybean genes differentially expressed under salt stress at 12 hr. The table includes mean expression values under control and stress conditions; fold change and $\log_2$ fold change values, *P*-values and adjusted *P*-values, and residual variance quotients of control and treatment samples. See Material and Methods for the criteria of differential expression.

Additional file 28: Table S8. Summary statistics of RNA-Seq analysis under dehydration and salt stress.

Additional file 29: Table S9. Raw read counts for each of the soybean gene under dehydration and salt stress at 0, 1, 6 and 12 hr generated in the RNA-Seq experiment.

Additional file 30: Table S10. DESeq normalized read counts for each of the soybean gene under dehydration and salt stress at 0, 1, 6 and 12 hr generated in the RNA-Seq experiment.

Additional file 31: Table S11. List of GO biological process and molecular function terms significantly (corrected $P < 0.05$) overrepresented in differentially expressed genes under dehydration and salt stress.

Additional file 32: Table S12. List of transcription factor classes significantly (corrected $P < 0.05$) overrepresented in differentially expressed genes under dehydration and salt stress.

Additional file 33: Table S13. List of plant transcription factor binding sites (TFBSs) significantly ($P < 0.05$, motif score >5) overrepresented in the promoters of HD-Zip genes,

differentially expressed (DE) genes under dehydration and salt stress, and their respective counts. The TFBSs are provided separately for each of the HD-Zip gene and DE genes.

Additional file 34: Table S14. List of plant transcription factor binding sites (TFBSs) significantly (P <0.05, motif score >5) overrepresented in the promoters of differentially expressed genes under dehydration and salt stress, with relative proportion of each TFBS under each of the stress treatment. The list of promoters that were excluded from the analysis because they did not meet the selection criteria (see material and methods for selection criteria) is included.

CHAPTER 5. CONCLUSIONS

Improvements in sequencing technologies have dramatically reduced sequencing costs. For example, the cost of sequencing a human-sized genome dropped from ~$95 M in September 2001 to ~$5 K in July 2014 (National Human Genome Research Institute). A major shift occurred in 2008 with the advent of second generation sequencing technologies – often referred to as next generation sequencing (NGS). NGS became a boon for both crops with and without any molecular resources (e.g. DNA or RNA sequences, expression datasets, or molecular markers such as "simple sequence repeats" (SSRs) and "single nucleotide polymorphisms" (SNPs), etc.). Before the advent of NGS, exploring the genetics of underutilized crops was challenging. Genetic characterization of a germplasm collection was performed with a few molecular markers. Some of the preferred marker technologies were "Randomly Amplified Polymorphic DNA" (RAPD) or "Amplified fragment length polymorphisms" (AFLPs), which did not require knowledge of sequence information. Additionally, performing genome-wide expression analysis was nearly impossible for underutilized crops. On the contrary, for crops with genome sequences, genome-wide expression analysis was possible using microarrays – but this was restricted to previously sequenced genes. NGS technologies, and specifically RNA-Seq, overcame most of these barriers. For example, in this dissertation, RNA-Seq assisted in the genomic resource development and genetic characterization of a collection of *Apios americana*, an underutilized crop with no prior molecular resources (Chapter 3), and also facilitated characterization of the soybean HD-Zip transcription factor family and identification of candidate genes involved in responses to dehydration and salt stresses (Chapter 4).

Mining the information in plant collections is the first step in cultivar development. This dissertation research has provided a first insight into morphology and genetics of the 53 remaining genotypes from Blackmon and Reynolds's Apios collection. Phenotypic characterization of the collection across multiple years, in multiple environments, and in different growing conditions has resulted in robust morphological descriptors for each of the genotypes (Chapter 2). Genetic characterization of the collection has provided a genetic fingerprint for each of the genotypes (Chapter 3). The phenotype data combined with the genetic information is now facilitating precise selection of parents in order to make crosses with the goal of developing high-yielding and "ideotype" cultivars.

Key steps in further improving Apios include: (1) increased public awareness; (2) creating markets - Apios may have a niche in flood-prone regions, and areas where protein malnutrition is widespread. It can sustain flooding, and yet produce protein rich tubers. Providing high yielding genotypes to smallholder farmers in different countries, where flooding and/or protein malnourishment exists, may help in creating markets; (3) supplementing Apios tubers to existing tuber and root crops - people are increasingly conscious about a healthy and nutritious diet. Apios tubers – with high protein content, isoflavones, and other likely-beneficial compounds – is an attractive supplement to existing tuber and root crops. Apios is cultivated in Japan and South Korea, because of its nutritional value relative to e.g. lower-protein potatoes and sweet potatoes. Healthy protein rich "French fries" and "fried chips" from Apios tubers make exciting delicacies; (4) developing ideotype cultivars - harvesting tubers requires considerable digging of the ground. Cultivars with shorter stolons, tubers spaced closely on the stolons, and are also high yielding will promote large-scale cultivation, and will make harvesting easier using the more efficient farm

machinery. The phenotype data combined with the genetic data has provided clues to select parents that may be crossed to get such desired cultivars (Chapter 3); (5) developing cultivars with determinate growth habit and high yield - Apios plants require trellising, and produce tubers that are often small. A cultivar development program is required to develop cultivars with increased yield and child-tuber size, decreased stolon length, and decreased vining tendency aboveground.

In conclusion, NGS – and specifically RNA-Seq – has facilitated rapid exploration of the underutilized crop *Apios americana*; and has allowed for comprehensive characterization of the soybean HD-Zip transcription factor family involved in abiotic stress responses, and identification of candidate genes for dehydration and salt stresses.

APPENDIX LICENSE INFORMATION

Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor family in soybean (*Glycine max*) during dehydration and salt stress

Vikas Belamkar, Nathan T Weeks, Arvind K Bharti, Andrew D Farmer, Michelle A Graham and Steven B Cannon

BMC Genomics 2014, 15:950 doi:10.1186/1471-2164-15-950

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2164/15/950

Evaluation of phenotypic variation in a collection of *Apios americana*: An edible tuberous legume

Vikas Belamkar, Alex Wenger, Scott R. Kalberer, V. Gautam Bhattacharya, William J. Blackmon and Steven B. Cannon

Reprinted by Permission, ASA, CSSA, SSSA

Crop Science 2015, 55(2):712-726 doi:10.2135/cropsci2014.04.0281

The electronic version of this article is the complete one and can be found online at: https://www.crops.org/publications/cs/articles/55/2/712

VITA

NAME OF AUTHOR: Vikas Belamkar
PLACE OF BIRTH: Bangalore, Karnataka, India

DEGREES AWARDED:

Bachelor of Engineering (B.E.) in Biotechnology, Sir M. Visvesvaraya Institute of Technology (Sir MVIT), Bangalore, India, 2007

M.S. in Biotechnology, Texas Tech University, 2010

HONORS AND AWARDS:

Iowa State University Research Excellence Award, 2014

Second place-poster competition, Plant and Animal Genome XXIII, San Diego, CA, 2015

Second place-poster competition, R.F. Baker Plant Breeding Symposium, Ames, IA, 2015

Best alumni successfully pursuing career in Biotechnology - Sir MVIT, Bangalore, 2015

PROFESSIONAL EXPERIENCE:

Research Assistant, Texas Tech University, 2007-2010

Research intern, Dow AgroSciences, Indianapolis, IN, May 20 to August 16, 2013

Research Assistant, Iowa State University, 2010-2015

INVITED TALKS

Translational genomics workshop, Plant and Animal Genome XXIII, San Diego, CA, 2015

VI International conference on legume genetics and genomics, Hyderabad, India, 2012

Plants for the future symposium, University of Missouri, Columbia, 2012

PROFESSIONAL PUBLICATIONS:

Vikas Belamkar, Alex Wenger, Scott R. Kalberer, V. Gautam Bhattacharya, William J. Blackmon and Steven B. Cannon. Evaluation of phenotypic variation in a collection of *Apios americana*: An edible tuberous legume. Crop Science 2015, 55:712-726.

Vikas Belamkar, Nathan T. Weeks, Arvind K. Bharti, Andrew D. Farmer, Michelle A. Graham and Steven B. Cannon. Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor family in soybean (*Glycine max*) during dehydration and salt stress. BMC Genomics 2014, 15:950.

Vikas Belamkar, Michael Gomez Selvaraj, Jamie L. Ayers, Paxton R. Payton, Naveen Puppala and Mark D. Burow. A first insight into population structure and linkage disequilibrium in the US peanut minicore collection. Genetica 2011, 139: 411-429.

Michael Gomez Selvaraj, Gloria Burow, John J. Burke, **Vikas Belamkar**, Naveen Puppala and Mark D. Burow. Heat Stress screening of peanut (*Arachis hypogaea* L.) seedlings for acquired thermotolerance. Plant Growth Regulation 2011, 65: 83-91.

N. Sathyanarayana, **P. B. Vikas**, Bharath Kumar T. N. and R. Rajesha. RAPD markers for genetic characterization in *Mucuna* species. Indian Journal of Genetics and Plant Breeding 2010, 70(3): 296-298.

N. Sathyanarayana, **P. B. Vikas**, R. Rajesha and T. N. Bharath Kumar. In vitro mass multiplication of *Abrus precatorius* (Linn.) through axillary bud culture. Journal of Cytology and Genetics 2008, 9: 57-63.

N. Sathyanarayana, R. Rajesha, **P. B. Vikas** and T. N. Bharath Kumar. Somatic embryogenesis & plant regeneration from stem explants of *Leptadenia reticulata Wight. & Arn*. Indian Journal of Biotechnology 2008, 7: 250-254.

N. Sathyanarayana, T. N. Bharath Kumar, **P. B. Vikas** and R. Rajesha. In vitro clonal propagation of *Mucuna pruriens* var. *utilis* and its evaluation of genetic stability through RAPD markers. African Journal of Biotechnology 2008, 7 (8): 973-980.