EXPLORING THE GENOMIC BASIS OF TRAITS RELEVANT TO EVOLUTION AND ECOLOGY OF CHESTNUT (*CASTANEA*) USING HIGH-THROUGHPUT DNA SEQUENCING AND BIOINFORMATICS

by

Nicholas R LaBonte

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Forestry and Natural Resources
West Lafayette, Indiana
December 2017

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Keith Woeste, Chair

USDA-FS Hardwood Tree Improvement and Regeneration Center

Dr. Robert Swihart

Department of Agronomy

Dr. Torbert Rocheford

Department of Agronomy

Dr. C. Dana Nelson

USDA-FS Southern Research Station

Approved by:

Dr. Robert Wagner

Head of the Graduate Program

ACKNOWLEDGMENTS

I would like to acknowledge the help of several people without whom the research described in this dissertation would not have been possible. Amy and Greg Miller helped introduce me to the world of chestnuts and helped to point me towards interesting avenues of research. I have the deepest gratitude towards the Northern Nut Growers Association (NNGA), American Chestnut Foundation (TACF) and the Indiana Academy of Sciences for the grants they provided; their financial support was essential to the project's success. For help in the lab and help procuring samples for research, I would like to thank Jim McKenna, Dr. Greg Miller (again), Dr. Peng Zhao, Dr. Laura Georgi, Ron Rathfon and Jason Tower of the Southern Indiana Purdue Agricultural Center (SIPAC), Bruce Wakeland, Claude Diehl, Aziz Ebrahimi, Margaret Woeste, Maha Ali, Marcia Kremer, Dr. Nate Lichti, and Dr. Jacqueline Doyle. Conversations with Mekala Sundaram were immensely helpful for interpreting the results of Chapter 4, so I would like to thank her. I thank Dr. Charles Michler for helping bring me to Purdue for my Master's thesis, and for helping convince to stay as a Van Eck Scholar to complete my Ph.D. I am grateful, and indebted, to Nathaniel Cannon and Dr. John Carlson of Penn State University for providing access to an improved Chinese chestnut genome assembly to carry out my work. I am indebted to my advisor, Dr. Keith Woeste, for his mentorship and guidance in completing this project, along with my committee members, Dr. Torbert Rocheford, Dr. C. Dana Nelson, and Dr. Robert Swihart, who generously lent their time and expertise to my educational endeavor. Finally, I would like to thank Dr. Cecilia Hennessy LaBonte, Theodora LaBonte, and Mathias LaBonte for their patience and support during the long process of finishing this great project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: LaBonte, Nicholas, R. PhD

Institution: Purdue University

Degree Received: December 2017

Title: Exploring the Genomic Basis of Traits Relevant to Evolution and Ecology of Chestnuts (*Castanea*) Using High-Throughput DNA Sequencing and Bioinformatics.

Committee Chair: Keith Woeste

Introduced pests and pathogens have devastated forest ecosystems in the temperate zone; in eastern North America, introduced pests and pathogens have led to the elimination of most mature elms (*Ulmus*), ashes (*Fraxinus*), hemlocks (*Tsuga*) and chestnuts (*Castanea*) over large areas where these genera were formerly abundant and important for local ecosystems. The restoration of species affected by introduced pests and pathogens requires the development and propagation of trees that possess heritable resistance. High-throughput DNA sequencing and genomics provide opportunities for researchers to identify resistance gene candidates, screen germplasm, and develop markers for marker-assisted selection in breeding programs, with the goal of restoring ecologically important wild trees to the landscape. American chestnut (*Castanea dentata*) is currently the focus of a major research effort that intends to restore the species by incorporating blight resistance from Chinese chestnut (*Castanea mollissima*), a species that is generally resistant to chestnut blight.

I investigated several aspects of chestnut genomics and blight resistance with the goal of aiding the blight resistance breeding program for American chestnut. I tested a detached-leaf assay for chestnut blight resistance and learned that it may not be useful for screening advanced backcross (BC3) progeny in chestnut blight resistance breeding programs (Chapter 2). Utilizing a recent draft assembly of the Chinese chestnut reference genome, I analyzed patterns of genetic variation across regions associated with chestnut blight resistance, and found that several loci associated with blight resistance show markedly elevated nucleotide diversity in the most resistant Chinese chestnuts relative to

more susceptible trees. At other blight-associated loci, genetic diversity was low in all *C. mollissima* (Chapter 3). This indicates that while maintaining high allelic diversity at blight resistance loci is desirable for a resistance breeding program, it may not be essential. Assessing potential unintended effects of hybrid breeding on the ecological behavior of restored chestnuts, I found that several genetic loci in third backcross (BC3) chestnut appear to affect caching decisions by squirrels due to inheritance of *C. mollissima* alleles that influence seed traits (Chapter 4). The reason for backcrossing in the American chestnut breeding program is to avoid the short, branchy mature form of *C. mollissima*. By sequencing the genomes of wild and orchard-derived Chinese chestnuts, I showed that some genomic loci under selection in orchard chestnuts (i.e., artificially selected by humans) may influence crown form (Chapter 5). This work should provide the basis for further investigations that validate the phenotypic effects of the proposed candidate genes, and utilize information on genetic polymorphisms identified here to accelerate chestnut improvement programs.

CHAPTER 1. OVERVIEW OF CHESTNUT

1.1 The Genus *Castanea*

The genus *Castanea* consists of 7 species of deciduous woody plants native to temperate forests in the Northern Hemisphere (Lang et al. 2007). East Asia contains the largest number of *Castanea* species: the Chinese chestnut (*Castanea mollissima*), Japanese chestnut (*C. crenata*), Chinese chinquapin (*C. henryi*), and Seguin chestnut (*C. seguinii).* Two species are found in North America, the American chestnut (*Castanea dentata*) and the Allegheny chinquapin (*C. pumila*), with a third (*C. ozarkensis*) sometimes recognized, although this tree is most often classified as a subspecies of *C. pumila*. In Europe and far western Asia, one species is native, the sweet chestnut (*Castanea sativa*). Based on genetic evidence (Lang et al. 2007), the center of diversity and point of origin for *Castanea* is in southeast Asia, perhaps around 60 million years before present (BP). *Castanea mollissima*, *C. henryi* and *C. seguinii* are each others' closest relatives, as are the North American *C. dentata* and *C. pumila*. Apparently, the one-nut-per-burr trait of *C. henryi* and *C. pumila* evolved separately in North America and Asia (Lang et al. 2007). By the Oligocene (30-25 million years BP) fossils similar to modern chestnuts appear in North America, in some areas still occupied by chestnuts (Tennessee) and some where they are no longer found (the Rocky Mountains). It is theorized that chestnuts dispersed from Asia to Europe, and then to North America during the middle Eocene (40 million years BP), founding the lineages that gave rise to today's species after continental drift divided the Tertiary deciduous forests (Lang et al. 2007).

Chestnuts belong to the Fagaceae family, along with oaks (*Quercus*), beeches (*Fagus*), and several smaller genera. Species in *Castanea* have simple, toothed leaves with an alternate arrangement. Flowers mature in late spring or early summer, with separate male and female inflorescences borne on the same tree. Self-pollination does not occur in chestnut, and pollination is mediated by wind, like other Fagaceae trees, but the numerous pollen-feeding insects that visit the large, conspicuous and scented male inflorescence (catkin) are also likely to carry out pollination. All chestnut species are diploid with 12 chromosomes (2n = 2x = 24). The genome is medium sized, between

780 and 800 million base pairs (Jacobs et al. 2012). As is common in outcrossing forest trees, chestnuts tend to be highly heterozygous, and high levels of gene flow appear to be maintained among wild populations (Worthen et al. 2010). Interspecific hybridization is possible among all species of chestnut, although partial infertility of F1s may result (Pereira-Lorenza et al. 2016).

Chestnuts usually constitute a minor component of mixed forests dominated by oaks; although they occasionally dominate forest stands, this is frequently the result of deliberate promotion of chestnuts by humans as a nut crop (in Europe and East Asia) or the relatively rapid regrowth of chesntut coppice following forest clearance (in North America). In general, chestnuts prefer high-light conditions for growth and regeneration, but can tolerate some shade, especially as seedlings. Their long lifespans (several centuries to over 1000 years) and tendency to resprout vigorously from the roots when felled make chestnuts a particularly durable component of the forests where they occur. On good sites, most chestnut species, with the exception of *Castanea pumila var. pumila,* can become large trees. *Castanea mollissima*, *C. crenata* and *C. sativa* are most frequently observed in orchard or open-forest environments where they grow as medium-sized, spreading trees. Like oak and beech, chestnuts produce large, nutritious seeds that are dispersed by birds and mammals that scatter-hoard seeds for later consumption. Nuts ripen inside spiny cupules (burrs), which hold 1, 3 or more nuts depending on the species. The burrs open in early autumn and release the seeds. The shells of chestnuts are thin, and the starchy, carbohydrate-rich nuts are easily consumed by a wide range of animals, including rodents, game birds, and ungulates. Tree squirrels (genus *Sciurus*), burying nuts soon after they fall, seem to be the most important dispersal agent of chestnut seeds, which desiccate quickly if left exposed on the ground.

Chestnuts are edible to humans with minimal processing, and lack the bitter taste (caused primarily by tannins) of most other nuts in the Fagaceae. For this reason, chestnuts are an important source of human nutrition in regions where they grow, particularly in China, the Korean peninsula, and the northern Mediterranean. In China, the cultivation of *Castanea mollissima* has occurred for thousands of years. In Europe, cultivation of *Castanea sativa* dates perhaps as far back as c. 3700 years BP based on pollen records (Rutter et al. 1991). Chestnut was widely utilized and moved around

Europe by the Romans, but may have been primarily introduced for coppice timber rather than nuts (Conedera et al. 2004). In the Middle Ages, orchards of grafted nut cultivars are well-attested in Italy. *Castanea crenata* is also cultivated as a nut crop in Japan and Korea. None of the American species were ever widely cultivated by indigenous American peoples, but certainly served as an important seasonal food source for people who lived among them, both before and after European settlement of the eastern United States. Today, chestnuts remain a popular food in Europe and East Asia. Chestnut production in North America, based mainly on Chinese chestnut but also some *C. crenata* x *sativa* hybrids, is increasing (Metaxas 2013).

1.2 American chestnut (*Castanea dentata*)

"But if a king is wholly vanished from our scene, its absence is at least less depressing than were those years when its diseased hosts and gaunt, whitening skeletons saddened the forest prospect."
-Donald Culross Peattie, A Natural History ofTrees of Eastern and Central North America, 1949

American chestnut (*Castanea dentata*) is the larger of the two North American chestnuts, and perhaps attained the largest dimensions of any chestnut species in the fertile coves of the southern Appalachian Mountains, reaching heights of 100-120 feet and 5-6 feet in diameter. Its closest relative is the Allegheny chinkapin (*C. pumila*), and it is more closely related to European chestnut than any of the Asian species (Jacobs et al. 2012). *Castanea dentata* has a native range extending north to Maine from the Appalachian foothills of Georgia and Alabama, and westward from the edge of the coastal plain to a few scattered sites in the hills of southern Illinois and Indiana and along the Mississippi River valley in Tennessee and Mississippi. From the pollen record, it seems that chestnut in the eastern U.S. underwent several fluctuations in abundance between the end of the last ice age and the present. From the historical record (1700s-present) it seems chestnut was generally a common but minor component of the forests where it occurred, except along the ridges of the southern Appalachians, where it made up a majority of stems in some areas, and in New England, where it was abundant. Some of this abundance may have been promoted by logging, as chestnut sprouts from stumps more consistently than oak and hickory. Chestnut was most often associated with these species, but across its wide range it appeared in a wide range of forest types- from mixed cove hardwood forests of the southern Appalachians to the sub-boreal pine/hardwood forests of Maine (Jacobs 2007).

In forests where it occurred, *Castanea dentata* was an ecologically important species because it produced a large volume of high-energy nuts with less annual variability in mast volume than the oaks (Dalgleish and Swihart 2012). This annual abundance of quality food for animals of all sizes augmented the entire food chain, meriting the American chestnut's designation as a keystone species. After Euro-American settlement of the Appalachians, wild chestnuts helped sustain rural human communities. Chestnuts were eaten, gathered and sold to urban markets, and used to fatten livestock before winter (Baxter 2009). The relatively light, decay-resistant wood was widely used for construction of fences, barns, and houses, and the tannin-rich bark was harvested to use in leather production. Salvageable chestnut wood is obtained from barns and even fences after a century or more of exposure to the elements in the damp, hot climate of the southeastern United States (Paillet 2002).

Human activity brought about the demise of the American chestnut as a keystone species not through purposeful exploitation, but rather through a series of biological accidents. At some point in the 1800s, the oomycete pathogen *Phytophthora cinnamomi* became established in the southeastern United States. Like many *Phytophthora* species this organism is an opportunistic pathogen, causing fatal root rot diseases in a wide range of susceptible plants, including *Castanea dentata* and *C. sativa,* the European chestnut. It is a native of Southeast Asia, and was most likely introduced either in nursery stock or soil used for ballast. Since it requires consistently warm, moist conditions to thrive, *P. cinnamomi* did not severely damage chestnuts in the Appalachians and New England, but may have eliminated many chestnut stands at lower elevations in the Southeast (Anagnostakis 2001).

Chestnut blight disease, caused by an ascomycete fungus that primarily attacks chestnut, was introduced in the early 1900s on trees imported from Japan and first observed in New York City in 1905 (Anagnostakis 1987). *Cryphonectria parasitica,* formerly *Endothia parasitica*, is a sexually-reproducing fungus that attacks chestnut cambium and phloem tissue, attacking through wounds or natural fissures in the bark. It produces a necrotic lesion (canker) of dead bark, through with orange fruiting bodies (the stroma) erupt in early summer. In resistant trees, cankers are surrounded quickly with vigorous callus tissue and contained. In susceptible trees, stems are typically girdled by a rapidly-expanding canker before callus forms, and if it does form, the *Cryphonectria* mycelium and the necrosis that follows usually expands around or over the callus tissue. Often, all tissue distal to the canker dies within one year. Unlike *Phytophthora,* which depends on wet soil for dispersal, the spores of *Cryphonectria parasitica* are wind-borne, and the pathogen thrives throughout *Castanea dentata's* native range. In spite of desperate early efforts to contain the pathogen, including a massive chestnut removal effort in Pennsylvania, it rapidly spread and top-killed nearly every American chestnut in the native range. By the 1930s, it had devastated the entire native range, effectively nullifying the ecological and human importance of American chestnut. The species survives today primarily as stump sprouts, although a few larger individuals survive around the fringes of, and outside of, the native range.

1.3 Chinese chestnut (*Castanea mollissima*)

Castanea mollissima, known in English as Chinese chestnut and in Chinese as *ban li*, has a large native range and produces a larger portion of the world's commercial chestnut harvest than any other species (Rutter et al. 1991). It is most often a relatively small tree, typically reaching about 40 feet in height in good conditions, and tends to have a branchy form with many large branches originating near the ground, although larger trees have been observed in remote forested areas. The size of nuts of Chinese chestnut is highly variable, with regional and cultivar differences apparent. In general, most Chinese chestnuts are larger than American chestnuts, but much smaller than the larger European or Japanese nuts. It has been cultivated in China for many centuries, and there is archaeological evidence for the utilization of wild chestnuts in Chinese prehistory, up to 6000 years ago (Jiangsu 1979). Chinese chestnut appears to be a native host of chestnut blight, as most trees are resistant to the disease and a few seem to be immune (Anagonstakis 1987). This has led to the use of *C. mollissima* as a blight resistance donor in hybrid chestnut breeding programs (with *C. dentata*) in the United States; in Europe, blight mortality is not severe enough to make such a breeding program necessary due to a hypovirus that causes attenuation of virulence in the blight fungus.

Castanea mollissima is mostly attested as an orchard tree: although wild populations exist (Fei et al. 2012), they are mostly located in remote areas and detailed accounts of silvics and ecology of wild *C. mollissima* continue to elude Western researchers. Most Chinese literature on chestnuts concerns orchard production, processing, and nutritional value, a logical focus given the large chestnut industry present in the country. One Chinese document, a technical handbook that was translated into English, distinguishes the "high and big" cultivated Chinese chestnut from another, "small" wild variety of *C. mollissima*, and notes that both these varieties and *C. seguinii* are often confused in some regions (Jiangsu 1979). Early germplasm collections (late 1800s and early 1900s) made by American scientists in China inevitably came from orchards because the remote, mountainous regions of China where wild chestnut forests occur were unsafe for travel at that time (Rutter 2004). Further explorations made by the American Chestnut Foundation (in 2011) identified some *C. mollissima* that seemed unlikely to be escapes from orchard cultivation, but most of the large forest trees found on this expedition were *C. henryi* and *C. seguinii* (Hirsh 2012). Several recent studies from the Chinese literature (e.g. Cheng et al. 2012) include samples from wild populations of *C. mollissima*, but it is not made clear how wild populations were distinguished from naturalized orchard-derived trees; most likely, proximity to orchards and human settelments was the determinant. In general, these studies have found that high genetic diversity, and low population differentiation, characterizes wild- and orchard-derived *C. mollissima* (Pereira-Lorenzo et al. 2016).

The native range of *C. mollissima* overlaps the smaller ranges of its congeners in China (Fei et al. 2012): presumably, there is some difference in ecological niche among the three species. The range of *C. mollissima* includes most of the Yangtze River valley, the mountains (Qinling and other ranges) that arc from near Beijing in the northeast to the edge of the Tibetan plateau in the southwest, and many areas in between; essentially, most of China except Manchuria, the deserts and plains of the west and Inner Mongolia, the highest mountains of Yunnan, and the tropical southern coast adjacent to Taiwan. It is also found in the Korean peninsula, although *C. crenata* is the chief orchard species there (Rutter et al. 1991). It is unknown how much of the native range of *C. mollissima* is due to planting and dispersal by humans, but it is likely that its distribution was less extensive prior to its adoption as a food plant (Fei et al. 2012). In ecological studies, Chinese chestnut tends to be found, like American chestnut, as a minor constituent in a wide variety of oak-dominated mixed forests, often in the understory. In northern China, it is sometimes found in wild stands with walnut (*Juglans regia*) and apricot *(Prunus armeniaca*) (Reisner 1921). Unsurprisingly, these stands are heavily utilized and modified by humans, and illustrate that *C. mollissima* is perhaps a difficult species to understand for American researchers because it most often occurs in a gray area between forest and orchard that does not exist in the United States. People living adjacent to natural chestnut-containing forests may manipulate the forest structure over time to favor chestnuts for nut production; while an orchard is never deliberately planted, the chestnut-dominated stand may ultimately function as an orchard.

Cultivated Chinese chestnuts can be divided into several different variety groups based on different centers of chestnut cultivation (Jiangsu 1978). The main center of commercial chestnut cultivation encompasses the range of the northern group, centered on Hebei Province (near Beijing), which is characterized by the smaller nuts generally preferred by Chinese consumers. Larger nuts are found in the Changjiang River Valley group in east-central China. Varieties from southern China are considered to be of lower quality in general, at least according to one report (Jiangsu 1978). In Northeastern China, near the Korean border, some *C. crenata* is apparently grown. The extent of hybridization between *C. mollissima* and *C. crenata*, where the latter occurs on the Asian mainland, is unknown; whether or not C. *mollissima* has any history of hybridization with the more closely related *C. seguinii* and *C. henryi* within the large area where the three species are sympatric is also unknown. Most *C. molllissima* material in the United States is believed to originate from coastal southern China (Rutter et al. 1991).

1.4 Interspecific hybrid breeding: history and status

Crossing between chestnut species, prior to the chestnut blight, was used to produce nut-producing varieties with some combination of valued parental traits (i.e., the size of

European chestnut with the flavor of American chestnut). Shortly after the start of the chestnut blight epidemic, a USDA breeding program was begun with the goal of deriving blight-resistant hybrids of American and Asian species to replace the lost forests of American chestnut (Clapper 1954, Berry 1978). Dr. Walter Van Fleet made thousands of interspecific chestnut crosses in the early 1900s in New Jersey; breeding programs led by Arthur Graves in Connecticut and R.B. Clapper in Maryland made and evaluated more crosses using several importations of Asian chestnut germplasm (Anagnostakis 2014). A smaller number of backcrosses were made to American and Asian parent species (Gravatt et al. 1954). In the 1960s, this breeding effort largely ceased after no blight-resistant trees were developed that combined strong blight resistance and the form of the American chestnut (Berry 1978), although S. Anagnostakis carried on the Connecticut-based program of Graves. The main problem was that Chinese chestnut's relatively small mature size and branchy form made it non-competitive in North American forests, and interspecific hybrids tended to inherit this crown form (Burnham et al. 1986). It is unclear whether the *C. mollissima* crown architecture, similar to that of an orchard-grown fruit tree and divergent from the "timber-type" form of American chestnut, is actually a domestication phenotype or is a result of different selective pressures in the forests of China, where canopy height is often lower (~50 feet) than in the native range of American chestnut (G. Miller, pers. comm.). Whatever its genetic basis and origin, this characteristic of *C. mollissima* and *C. crenata* led to the cessation of early breeding programs when no "timber-type" blight-resistant hybrids were produced. *C. seguinii* and *C. henryi* may not have had this problem, but these species did not perform well when planted in the United States (Berry 1978).

In the 1980s, the idea of incorporating resistance genes from Chinese chestnut into an American chestnut was revived by Dr. Charles Burnham and associates (Burnham et al. 1986). Burnham's background in row crop breeding made him familiar with backcross breeding, which he proposed as a way to recover a "timber type" tree from a hybrid of *C. dentata* and *C. mollissima.* By crossing resistant hybrids back to American chestnut for three generations, selecting only resistant trees with good form at each stage, a tree resembling *Castanea dentata*, but possessing the blight resistance genes of *C. mollissima*, could be derived. Importantly, it was theorized that two incompletely dominant resistance loci controlled blight resistance in chestnut: backcross breeding is not effective when a large number of genes are transferred. The American Chestnut Foundation (TACF) was formed to carry out the backcross breeding scheme. Two first-backcross (BC1; (*C. dentata x C. mollissima*) x *C. dentata*) individuals, nicknamed “Clapper” and “Graves,” were identified as exceptionally vigorous and blight-resistant in test plantings from the older USDA breeding program and selected as the resistant parents for the TACF breeding program to “jump-start” the process. State chapters of TACF crossed local *C. dentata* germplasm with the ‘Clapper’/’Graves’ hybrid material to establish locally adapted lines. Today, a large number of BC3F2 orchards (intercrosses of third-backcross (BC3) trees) have been established: these trees closely resemble American chestnut in most phenotypic characters. Once the most resistant trees have been selected from this population, the breeding program is supposed to enter its final stages. So far, recovery ofAmerican chestnut characteristics not related to blight resistance has been successful (Diskin et al. 2006). Blight-resistance is measurably improved in backcross populations relative to American chestnut, and TACF is incorporating more *C. mollissima* resistance donors in an effort to improve resistance further (Hebard 2005, 2006).

Genetics and genomics research on chestnuts in the United States has been conducted mainly to augment the TACF breeding program and spur on the restoration of American chestnut to the eastern U.S. One major research goal has been to identify the sites in the genome that make Chinese chestnuts more resistant to chestnut blight than American chestnut. Kubisiak et al. (1997) identified three major quantitative trait loci (QTL) that together account for around 75% of the variation in blight resistance in an intercross of interspecific hybrids. These loci were confirmed, in the same test population, using a larger and more sophisticated set of markers (Kubisiak et al. 2013). Selecting three loci in a backcross breeding scheme is more difficult than transferring 1 or 2 loci, but is still possible. A transcriptomics study comparing mRNA in *C. dentata* and *C. mollissima* tissues following blight canker initiation identified transcripts that were differentially expressed in infected stem tissue of American and Chinese chestnuts and furnished a high-quality transcriptome assembly for chestnut (Baraket et al. 2009, 2012). A draft reference genome exists for *C. mollissima*; the genetic map and physical map have been integrated (Fang et al. 2012) and the refinement of the draft genome as a high-quality reference sequence is ongoing (Staton et al. 2014). A set of genome scaffolds corresponding to the chromosome locations of the three main chestnut blight resistance QTL sequences has been identified (Staton et al. 2015).

As the TACF breeding program and research on chestnut genetics and genomes advances, opportunities present themselves to improve the blight resistance breeding program and investigate some important questions that lack adequate answers. In this dissertation, I detail my investigation of several questions related to the biology, evolution, and reintroduction of chestnut in North America.

1) Can backcrossed chestnuts be screened for blight resistance more efficiently using a detached-leaf assay, developed by Newhouse et al. (2014)?
2) How do the genomes of blight-resistant Chinese chestnuts, blight-susceptible Chinese chestnuts, and susceptible chestnut species differ across genomic regions associated with blight resistance?
3) How do the genomes of cultivated Chinese chestnuts differ from those of wild conspecifics?
4) What are the genetic factors that control differences in seed dispersal among interspecific chestnut hybrids?

1.5 Literature Cited

Anagnostakis SL (1987) Chestnut blight: the classical problem of an introduced pathogen. Mycologia 79(1): 23-27.

Anagnostakis SL (2001) The effect of multiple importations of pests and pathogens on a native tree. Biological Invasions 3: 245-254.

Anagnostakis SL (2014) Chestnut breeding in the United States. The Connecticut Agricultural Experiment Station, 03/05/2014.

Barakat A, DiLoreto, DS, Zhang Y, Smith C, Baier K, Powell WA, Wheeler N, Sederoff R, Carlson JE (2009) Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (C. mollissima) in response to the chestnut blight infection. BMC Plant Biology 9:51.

Barakat A, Staton M, Cheng C-H et al. (2012) Chestnut resistance to the blight disease: insights from transcriptome analaysis. BMC Plant Biology 12:38.

Baxter, BN (2009) An oral history of the American chestnut in southern Appalachia. Master’s thesis, University of Tennessee at Chattanooga, 109 pp.

Berry, FH (1978) Chestnut breeding in the United States Department of Agriculture. USDA Forest Service Northeastern Forest Experiment Station, Delaware, OH. 2 pp.

Burnham CR, Rutter PA, French DW (1986) Breeding Blight-Resistant Chestnuts. Plant Breeding Reviews 4: 347-397.

Cheng L-L, Feng H-D, Rao Q, Wu W, Zhou M, Hu G-L, Huang W-G (2012) Diversity of wild Chinese chestnut chloroplast DNA SSRs in Shiyan. J Fruit Sci 3:382-386.

Clapper RB (1954) Chestnut Breeding, Techniques and Results. I. Breeding material and pollination techniques. Journal of Heredity 45: 106-114, 201-218.

Conedera M, Krebs P, Tinner W, Pradella M, Torriani D (2004) The cultivation of *Castanea sativa* (Mill.) in Europe, from its origin to its diffusion on a continental scale. Vegetation History and Archaeology 13(3):161-179.

Dalgleish HJ, Swihart RK (2012) American chestnut past and future: implications of restoration for resource pulses and consumer populations of eastern U.S. forests. Restoration Ecology 20(4): 490-497.

Diskin M, Steiner KC, Hebard FV (2006) Recovery of American chestnut characterstics following hybridization and backcross breeding to restore blight-ravaged Castanea dentata. Forest Ecology and Management 223:439-447.

Fang G-C, Blackmon BP, Staton ME, Nelson CD, Kubisiak TL et al. (2012) A physical map of the Chinese chestnut genome and its integration with the genetic map. Tree Genetics and Genomes 9(2):525-537.

Fei S, Liang L, Paillet FL, Steiner KC, Fang J, Shen Z, Wang Z, Hebard FV (2012) Modeling chestnut biogeography for American chestnut restoration.

Gravatt GF, Diller JD, Berry FH, Graves AH, Nienstaedt H (1954) Breeding timber chestnuts for blight resistance. Proc. 1st NE Forest Tree Improvement Conferences: 70-75.

Hebard FV (2005) Meadowview Notes 2004-2005. Journal of the American Chestnut Foundation 19(2): 16-27.

Hirsch R (2012) Chestnuts in China: a science expedition turns into an adventure. Journal of the American Chestnut Foundation 6(26):22-25.

Hebard FV (2006) The backcross breeding program of the American Chestnut Foundation. pp. 61-77. In: K.C. Steiner and J.E. Carlson (eds.), Restoration of the American chestnut tree to forest lands- proceedings of a conference and workshop. May 4-6 2004, North Carolina Arboretum. Natural Resources Rep. NPS/NCR/CUE/NRR- 2006/001, National Park Service, Washington, D.C.

Huang H, Carey WA, Dane F, Norton JD (1996) Evaluation of Chinese chestnut cultivars for resistance to *Cryphonectria parasitica*. Plant Disease 80: 45-47.

Jacobs DF (2007) Toward development of silvical strategies for forest restoration of American chestnut (*Castanea dentata*) using blight-resistant hybrids. Biological Conservation 137(4):497-506.

Jacobs DF, Dalgleish HJ, Nelson CD (2012) A conceptual framework for restoration of threatened plants: the effective model of American chestnut (*Castanea dentata*) reintroduction. New Phytologist 197: 378-393.

Jiangsu Institute of Botanical Research (1979) Ban Li (Chestnut) Science Publishing House, Beijing, China. (In Chinese, partial translation available from The American Chestnut Foundation).

Kubisiak TL, Hebard FV, Nelson CD, Zhang J, Bernatzky R, Huang J, Anagnostakis SL, Doudrick RL (1997) Molecular mapping of resistance to blight in an interspecific cross in the genus *Castanea*. Phytopathology 87:751-759.

Kubisiak TL, Nelson CD, Staton ME, Zhebentyayeva T, Smith C, Olukolu BA, Fang G-C, Hebard FV, Anagnostakis S, Wheeler N, Sisco PH, Abbott AG, Sederoff RR (2013) A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*). Tree Genetics and Genomes 9:557-571.

Lang P, Dane F, Kubisiak TL, Huang H (2007) Molecular evidence for an Asian origin and a unique westward migration of species in the genus *Castanea* via Europe to North America. Molecular Phylogenetics and Evolution 43(1):49-59.

Metaxas A (2013) Chestnut (*Castanea* spp.) cultivar evaluation for commercial chestnut production in Hamilton County, Tennessee. M.S.E.S Thesis, University of Tennessee at Chattanooga, 135 pp.

Newhouse AE, Spitzer JE, Maynard CA, Powell WA (2014) Chestnut leaf inoculation as a rapid predictor of blight susceptibility. Plant Dis. 98(1):4-9.

Paillet FL (2002) Chestnut: History and ecology of a transformed species. Journal of Biogeography 29: 1517-1530.

Peattie DC (1950) Trees of Eastern and Central North America. Houghton Mifflin Company, Boston, MA: p. 119-121.

Pereira-Lorenzo S, Lourenço Costa R, Anagnostakis S et al. (2016) Interspecific hybridization of chestnut. In: Polyploidy and Hybridization for Crop Improvement, Chapter: 15, Publisher: CRC Press, Editor: Mason AS, pp. 379-408.

Reisner JH (1921) Nut culture in China. American Nut Journal 14(2):17.

Rutter, PA, Miller G, Payne JA (1991). Chestnuts (*Castanea*). ISHS Acta Horticulturae 290: Genetic Resources of Temperate Fruit and Nut Crops (761-788).

Staton ME, Addo-Quaye C, Cannon N, Tomsho LP, Drautz D, Wagner TK, Zembower N, Ficklin S, Saski C, Burhans R, Schuster SC, Abbott AG, Nelson CD, Hebard FV, Carlson JE (2014) *The Chinese chestnut (Castanea mollissima) genome version 1.1*, http://www.hardwoodgenomics.org/chinese-chestnut-genome, Access date August 2, 2016.

Staton ME, Zhebentyayeva T, Olukolu B, Fang GC, Nelson D, Carlson JE, Abbott AG (2015) Substantial genome synteny preservation among woody angiosperm species: comparative genomics of Chinese chestnut (*Castanea mollissima*) and plant reference genomes. BMC Genomics 16:744.

Worthen LM, Woeste KE, Michler CH (2010) Breeding American chestnuts for blight resistance. Plant Breeding Reviews 33: 305-339.

CHAPTER 2: EFFECTIVENESS OF A DETACHED LEAF ASSAY AS A PROXY FOR STEM INOCULATIONS IN BACKCROSSED CHESTNUT (*CASTANEA*) BLIGHT RESISTANCE BREEDING POPULATIONS.

Published in: Forest Pathology 47(2): e12313
April 2017

Authors: Nicholas LaBonte, James McKenna, Keith Woeste

2.1 Introduction

The backcross breeding program of the American Chestnut Foundation (TACF), which builds on the work of earlier hybrid breeding programs initiated during the chestnut blight epidemic of the early 20th century, aims for the restoration of chestnut to the forests of eastern North America (Gravatt et al. 1953; Burnham et al. 1986). First detected in 1905, chestnut blight, caused by the ascomycete *Cryphonectria parasitica (Cp),* spread rapidly throughout the native range of the American chestnut and eliminated it as a canopy species (Anagnostakis 1987). Chestnut blight causes necrotic cankers on the surface of the branches and trunk that can expand to cause girdling and death in susceptible trees. Since Chinese chestnut is the most resistant species to chestnut blight and readily hybridizes with American chestnut, it serves as the resistance donor for the breeding program. Evaluations of hybrid crosses led breeders to hypothesize that a few major genes control blight resistance, so a backcrossing program is a reasonable way to derive trees that look like American chestnut but are highly blight-resistant (Diskin et al. 2006). Quantitative trait locus (QTL) mapping studies to date have supported this hypothesis (Kubisiak et al. 1997, 2013). TACF has proposed a plan to backcross chestnuts for three generations, then intercross the third backcross progeny with high levels of resistance to produce a generation of progeny (B3F2) in which a few recombinant individuals are homozygous for all the major resistance genes. A large number of trees must be evaluated at this stage (1200 per family) because the desired recombinants are rare, and a large range of phenotypes from highly-susceptible to highly-resistant are present (Burnham et al. 1986, Fitzsimmons et al. 2014).

The traditional method for evaluating chestnut blight resistance is a stem inoculation made in early summer on four- or five-year-old trees with a cork borer and a small agar plug of inoculum (Hebard 2005). Cankers develop and are evaluated in late fall and/or the following summer. This method is reliable but requires large amounts of time and land to grow trees. A method for inoculating the small-diameter stems of first- or second-year chestnut trees (Powell et al. 2007) was never widely adopted. The leaf-inoculation method published by Newhouse et al. (2014) generated interest in the chestnut breeding community for several reasons- trees can be inoculated and highly susceptible trees removed in the first year, the test is not fatal to the tree being tested, and scoring is straightforward, rapid, and quantitative.

Detached-leaf assays are often used in woody plant breeding programs as rapid alternatives to field inoculations when such methods can be correlated with field inoculations and, ultimately, robust disease resistance reactions (e.g. Tahi et al. 2000, Calonnec et al. 2012). In tree breeding, reducing the time it takes to evaluate crosses has great appeal. However, detached-leaf assays are most commonly utilized for pathogen species that naturally attack leaf tissue (Calonnec et al 2012), although leaf assays have been developed for non-leaf pathogens with some success (Tedford et al. 1990, Francis et al. 2010). *Cp* has not been documented as a leaf pathogen. It is possible that tree defense mechanisms against pathogens that attack the stem and those that attack leaf surfaces are different enough that a leaf assay would be ineffective, but Newhouse et al. (2014) showed that their leaf lesion assay accurately discriminated among resistant Chinese chestnut, susceptible American chestnut, and a third species considered intermediate, the Allegheny chinquapin (*Castanea pumila*).

We saw an opportunity to test leaf inoculations on a large population of B3F2 chestnuts that had received stem inoculations for rating *Cp* resistance in the summer of 2013. Our objective was to use this population to determine if screening of B3F2 offspring using leaf inoculations could improve the efficiency of the Indiana state chapter TACF breeding program. Since major blight resistance genes are segregating in this population, we expected to observe stem and leaf inoculation phenotypes spanning most of the range of variability between Chinese chestnut and American chestnut. Our purpose was to extend the new detached-leaf assay to a practical breeding application by testing whether leaf inoculation could serve as a proxy for stem inoculations, and thereby allow susceptible trees to be rogued as seedlings in TACF's breeding program.

2.2 Materials and Methods

2.2.1 Test Populations

In 2014, 100 B3F2 trees were screened using the detached-leaf assay from each of two planted sites: one at the Southern Indiana Purdue Agricultural Center (SIPAC) in Dubois County and the other at the Potawatomi Wildlife Park in northern Indiana's Marshall County (PWP). Trees at each planted site represented a full-sib family (B3 × B3). A majority of the trees sampled at SIPAC had been tested using stem inoculations by late June 2014 (Table 1). In addition, in 2014, five Chinese chestnuts and five early-generation backcross trees (primarily B3) planted at the Purdue University Lugar Farm (LF), along with five American chestnuts at Purdue's Martell Forest (MF), both located in Tippecanoe County, Indiana were screened using leaf inoculations (Table 1). In 2015, trees from PWP were not re-screened using leaf inoculations, but surviving trees from SIPAC plus a set of younger trees at SIPAC (from the same full-sib family) that had received stem inoculations in early June 2015 (n=135) were sampled for leaf inoculations. Additionally, an expanded set of early-generation backcross trees that had been stem-inoculated in 2014 were screened at Purdue using the detached-leaf assay (n=49) in the summer of 2015 (Table 1).

2.2.2 Stem Inoculations

Stem inoculations were performed in June of 2013, 2014, and 2015. A cylinder of bark was removed with a 6 mm cork-borer and an agar plug containing either a highly aggressive *Cp* strain (Ep155) or a less-aggressive strain (Sg88) was inserted into the cambium and taped into place following standard TACF protocol (Griffin et al. 1983). Fungal cultures were obtained from Dr. Fred Hebard of TACF in Meadowview, VA. Cankers were rated on an ordinal qualitative scale in November and July following inoculation: 1 (small, tightly contained canker) to 5 (large, non-contained canker) (Hebard 2005). Length and width of developing cankers were also measured in

September 2014 (measurements of developing cankers from June 2014) and September 2015 (measurements of cankers developing cankers from June 2015).

2.2.3 Leaf Inoculations

We closely followed the protocol of Newhouse (2014). We sampled 5 to 8 leaves per tree for B3F1s and B3F2s and 10 leaves per tree for resistant and susceptible species controls, taking care to select leaves that were fully expanded but still tender, generally from the ends of shoots. In 2015, on trees that exhibited basal shoots and surviving live crown branches, four leaves were taken from first-year basal shoots and four from older crown branches, to test whether the different characteristics of these leaves had any effect on leaf inoculations. Leaves were rinsed in one bath of 0.1% Tween and two baths of distilled water, patted dry, labeled with permanent marker, inoculated with *Cp*, and stored for 5 to 6 days in Sterilite gasket-sealed plastic boxes of the same size and model as those used by Newhouse et al. (2014) that were lined with damp paper towels. Sealed plastic boxes were held in the dark at room temperature until symptoms were measured. In 2014, the strain sg88 was used; in 2015, ep155 and sg88 were placed on an equal number of leaves from each tree sampled. *Cp* was cultured on acidified potato dextrose agar; cultures were stored in the dark at room temperature until they had reached sufficient size (4-5 days) and inoculum was always taken from the expanding edge of the colony. Agarose plugs were balanced on a cut (about 5 mm in length) in the mid-vein of the abaxial side of the leaf, made with a razor blade (sterilized with 70% ethanol). Leaves were randomly assigned to boxes to avoid confounding of individual effects with any effects due to different conditions in the boxes. Lesion length and width were measured with a digital caliper, but lesion length alone was used for analysis because the two variables were correlated.

2.2.4 Statistical methods

We used Tukey’s multiple comparison tests in R (R Foundation for Statistical Computing 2015) to test for differences in the mean leaf lesion length of different species/ backcross generations of chestnut and B3F2 chestnuts grouped according to their rating of stem canker symptoms (1 to 5). ANOVA (R function: aov) was also used to analyze the variability of leaf lesion dimensions among chestnuts in different blight resistance classes. We used linear regression (R function: lm) to identify whether there was a relationship between stem canker length and leaf lesion length, to identify correlations between canker length and width, and assess correlations between years for an individual tree's leaf lesion scores. Because the stem and leaf canker measurements were quantitative and normally distributed (not shown), we were confident that the basic assumptions of parametric regression and analysis of variance were met.

2.3 Results

2.3.1 Leaf inoculation protocol

Lesion development on leaves from susceptible hosts conformed to the description in Newhouse et al. (2014). Control inoculations with blank agar plugs failed to produce lesions, and in cases where the agar plug rolled away from the inoculation site, no necrotic lesions were observed along the cut. Secondary infections away from the inoculation point were only observed on a few leaves, and inoculations rarely failed to produce visible lesions, although failures were more frequent on Chinese chestnut than on American or hybrid leaves. In 2014, 15 of 50 Chinese chestnut leaf inoculations failed to produce symptoms, while only 1 of 50 American chestnut leaves and 0 of 50 B3F1 leaves failed to develop lesions. Inoculations with sg88 resulted in leaf lesion sizes that were not significantly different than inoculations with ep155 based on 95% confidence interval estimates of the means for lesion length and width. The mean lesion width for ep155 was 13.25 +/- 0.75 mm compared to 12.76 +/- 0.82 mm for sg88; mean lesion length for ep155 was 28.69 +/- 1.43 mm, and for sg88 28.48 +/- 1.61 mm. In leaves of B3F2 trees from SIPAC, the difference between the two strains was similarly small (ep155 mean length = 28.51 mm; sg88 = 28.01 mm). For this reason, lesions from the different isolates were pooled for subsequent analysis. Leaf lesion length and width were correlated (Pearson's correlation coefficient: 0.79).

2.3.2 Variability in leaf lesion size by site, year, and species

A Tukey multiple-comparisons test for differences in mean leaf lesion length among B3F2 chestnuts at PWP, B3F2 at SIPAC, B3F1 at FNR, American chestnut at FNR, and

Chinese chestnut at FNR in 2014 revealed a significant difference between the mean of Chinese chestnut versus all other groups sampled, and between the PWP B3F2 planting versus all other groups, while the difference between SIPAC and American chestnut was not significant after correction for multiple tests (Table 2). The PWP samples developed the largest leaf lesions by far, perhaps because of nutrient deficiencies the trees suffered from there, which led to the exclusion of genotypes from this site from subsequent sampling and testing (Table 2). In 2015, when samples from SIPAC, Chinese chestnuts and B3F1 were included, we identified a significant difference in leaf lesion length between Chinese and American, B3F2, and B3F1 chestnuts, while the latter three groups did not differ significantly from each other. For each sampled genotype, lesion length was significantly correlated between 2014 and 2015 samples ($r^2 = 0.30$; $p < 0.001$), indicating a low to moderate reproducibility in length of leaf lesions across years. There were no apparent effects of year on leaf lesion length or width (overall mean length 2014 = 30.8 mm; 2015 = 29.4 mm; mean width 2014 = 12.37 mm; 2015 = 12.57 mm).

2.3.3 Relationship of leaf lesion length to canker rating

ANOVA tests and Tukey's multiple comparison of means tests were used to determine if leaf lesion size differed for genotypes of the SIPAC B3F2 trees from different resistance categories based on stem inoculations (1 to 5) (Figure 1). Of the four ANOVA tests performed, only one (2015 leaf lesion length by 2015 stem canker rating) indicated that there was any difference in leaf lesion dimensions among trees in the different stem canker rating categories ($F_{(1,87)} = 10.67$, $p = 0.001$). In this case, the Tukey HSD test supported a difference in mean leaf lesion length for trees in category 1 (most resistant based on stem lesion phenotype, mean leaf lesion length = 20.58 mm) versus genotypes with stem cankers rated 3, 4, or 5 (mean leaf lesion length 26.93 mm, 27.06 mm, 27.42 mm, respectively). All the trees with stem canker ratings of 2 in 2014 had stem canker ratings of 3, 4, or 5 by 2015.

2.3.4 Relationship of leaf lesion length to stem canker length

Simple linear regressions indicated that while the association between leaf lesion measurements and stem canker length was occasionally significantly different from zero

(Table 3), only very weak associations were observed, based on estimated regression coefficients and r^2 values. In some cases, estimated values for the slope of the regression of leaf and stem canker dimensions were negative. When the same analysis was performed on the B3F1 and Chinese trees that had received stem inoculations at the Lugar farm, no significant correlations between leaf lesion length and stem canker length were found (results not shown).

2.4 Discussion

We observed differences in leaf lesion size between *Cp*-resistant and *Cp*-susceptible chestnut species, as was observed in Newhouse et al. (2014). We also observed variation among individual B3F2s in leaf lesion size that was somewhat consistent between the two years of the study, but the most resistant and the most susceptible B3F2 trees based on stem inoculations had leaf lesions that were mostly indistinguishable in size, and the size of leaf lesions of B3F2s in general was similar to the mean size of leaf lesions of American chestnut. B3F2 genotypes would be expected to have a wide range of leaf lesion phenotypes if leaf lesions reflected overall resistance to *Cp* and if the model of inheritance of resistance to *Cp* that is assumed by the TACF breeding program is correct. It was also expected that if leaf lesions reflected overall resistance to *Cp* and if the model of inheritance of resistance to *Cp* that is assumed by the TACF breeding program is correct, then mean lesion length of inoculated leaves should be intermediate between lesion length of American and Chinese chestnuts, or at least distinguishable from the mean of American chestnut. In both 2014 and 2015, while clear differences between American and Chinese chestnut leaf lesion dimensions were observed, and B3F2s displayed a range of leaf lesion sizes, those sizes were not intermediate but closely matched values for American chestnut.

Why did our results not conform to expectations of the current chestnut blight resistance breeding model? One reason may have been the generally low blight resistance (based on stem inoculations) in the populations tested. A majority of B3F2 trees at SIPAC were rated highly susceptible (rated either 4 or 5) in both 2014 and 2015, neither of which was an unfavorable year for chestnut growth. The very small number of highly resistant B3F2 trees rated 1 or 2 (n=9 in 2014 and n=3 in 2015) had significantly smaller leaf lesion dimensions in 2015 (Figure 1), but there was no significant difference in leaf lesion size between moderately resistant (stem canker rating = 3) and highly susceptible trees (Figure 1). It is possible that the leaf lesion size of the most resistant trees was influenced by the greater vigor of those trees relative to moderately-susceptible individuals, or more likely, leaf lesion size was influenced by the morphology of the inoculated leaves. Some trees that rated highly resistant (1 or 2) for stem cankers and had small (2 standard deviations less than the mean) leaf lesions in 2014 deteriorated in both categories by the summer of 2015.All the B3F2 trees rated “1” in 2015 were stem inoculated in 2014, and all B3F2 trees stem-inoculated in 2013 showed symptoms rated > 2 by 2015. The observed increase in both stem canker rating and leaf lesion size from 2014 to 2015 among trees rated 1 and 2 for stem cankers may have been a reflection of relative susceptibility to *Cp*, a reflection of morbidity incited by *Cp*, or both, and the reasons for the increase may not have been the same for stem and leaf tissues. However, the few trees that maintained a low to moderate (1, 2, 3) canker severity rating in both 2014 and 2015 (n = 4) had relatively small leaf lesions: 2014 mean lesion length for these trees was 23.53; in 2015 it was 26.18 (compared to overall means of 30.8 and 29.4 for 2014 and 2015, respectively). Based on these results, it seems practical to use leaf inoculations to eliminate the most susceptible trees at a young age (e.g., 25% of trees with largest leaf lesions could be rogued). Eliminating the most susceptible trees based on leaf lesion size would be unlikely to lead to accidentally discarding the most resistant members of the B3F2 population, but given that leaf inoculations could not consistently distinguish moderately susceptible (possible desirable) trees from the most susceptible trees, the utility of this method to breeders is inferior to that of traditional stem inoculations.

Detached-leaf assays developed as proxies for inoculation of other tissues have not always been effective in cases where fruits (Liebhard et al. 2003) or roots (Irwin et al. 2003) were the pathogen’s target tissue, and their utility to assay stem pathogens in forest trees has been questionable in some other cases (Parke et al. 2005). The trees we inoculated and evaluated were selected because they presented an opportunity to test the published leaf inoculation method against the current standard for measuring susceptibility to *Cp*, stem inoculations. Since our tests took place on 5-year-old field-grown trees, there were some potential confounding factors that would probably not be present if screening was performed using greenhouse-grown seedlings. First, as mentioned above, there is the potential that the severity of a tree's reaction to stem inoculation affected the results of the leaf inoculation. In particular, we were concerned that very susceptible trees, which were killed above the inoculation point in the first year, might have biased results because the leaves being tested inevitably came from shoots below the inoculation point. Leaves from these shoots could have elevated levels of defense compounds that would affect the development of leaf lesions, but when we compared the mean canker size of shoot leaves with those from established branches, the lesions that developed on shoot leaves were, on average, slightly larger (29.1 mm vs. 28.5 mm). The sample for the comparison included shoot leaves from both highly susceptible and somewhat resistant trees. Therefore, any bias from the use of shoot leaves would have been in the direction of susceptible trees developing larger lesions. Site conditions also obviously had a large effect on leaf lesion size: leaves from B3F2 chestnuts at the sub-optimal PWP site developed even larger lesions than American chestnut leaves at Martell Forest. Tahi et al. (2000) described a case where a detached leaf assay deemed ineffective when tested with field-grown leaves later proved to be useful when greenhouse specimens were tested. It is likely that greenhouse-grown B3F2 seedlings would provide better material for leaf inoculations and a test using this approach (with seedlings tested in the greenhouse and evaluated using stem inoculations in the field 5 years later) could validate the method for hybrid breeding.

We hypothesize that the differences in leaf lesion size between susceptible American chestnut and resistant Chinese chestnut, observed by Newhouse et al. and replicated in this study, were caused not only by the defensive mechanisms that confer blight resistance to Chinese chestnut, but also by morphological and histological differences in the leaves of the species. Chinese chestnut has heavy, waxy leaves with a densely hairy underside, while American chestnut leaves are not hairy and are generally thinner and less heavily suberized. Since gross phenotypic characters of Chinese chestnut are selected against in the TACF breeding program, by three generations of backcrossing most Chinese-like leaf characteristics have been eliminated (Diskin et al. 2005). This could explain the general similarity of B3F2 reaction to the American chestnut reaction when leaves were inoculated, even when the B3F2 trees manifest some level of blight resistance from the Chinese donor parent.

Furthermore, the intermediate species used by Newhouse et al. when first describing the assay, Allegheny chinkapin, has some leaf characteristics in common with Chinese chestnut, namely, pubescence on the abaxial side of the leaf. Finally, two putative F1 or B1 (first-generation hybrid or first-backcross) trees were among those inoculated at Purdue's Lugar farm. These trees, which had hairy, waxy leaves intermediate between Chinese and American chestnut, had smaller leaf lesions than the B3 trees assayed at the same site. Unfortunately, these were the only early-generation hybrids we had access to for the study: a test of F2 hybrid trees with varying levels of Chinese-like leaf trait expression would be a good test of the hypothesis that leaf traits control the leaf phenotype in addition to inherent blight resistance.

Our study, inspired by the excitement generated by the potential of the detached-leaf among chestnut breeders, sought to extend the results of Newhouse et al. (2014) from comparisons of resistant and susceptible chestnut species to the backcrossed hybrid trees that TACF hopes to use for the restoration of chestnuts to the North American landscape. We conclude that the leaf-inoculation assay does not discriminate between resistant and susceptible trees under field-based breeding conditions, such as that conducted by IN-TACF. Further research to improve the utility of this assay should compare greenhouse-grown versus field-grown leaves and examine the effects of leaf morphological differences in greater depth.

2.5 Literature cited

Anagnostakis SL (1987) Chestnut blight: the classical problem of an introduced pathogen. Mycologia 79(1):23-37.

Burnham CR, Rutter PA , French DW (1986) Breeding blight-resistant chestnuts. Plant Breed. Rev. 4:347-397.

Calonnec A, Wiedemann-Merdinoglu S, Deliere L, Cartolaro P, Schneider C, Delmotte F (2012) The reliability of leaf bioassays for predicting disease resistance on fruit: A case study on grapevine resistance to downy and PWPdery mildew. Plant Pathol. 2012:1-13.

Diskin M, Steiner KC, Hebard FV (2006) Recovery of American chestnut characteristics following hybridization and backcross breeding to restore blight-ravaged *Castanea dentata*. For. Ecol. and Manage. 223:439-447.

Fitzsimmons S, Gurney K, Georgi L, Hebard F, Brinckman M, Saielli (2014) Regionally adapted seed orchards within TACF's state chapters. Journal of the American Chestnut Foundation 28(1):15-19.

Francis MI, Peña A, Graham JH (2010) Detached leaf inoculation of germplasm for rapid screening of resistance to citrus canker and citrus bacterial spot. Eur. J. Plant Pathol. 127(4):571-578.

Gravatt GF, Diller JD, Berry FH, Graves AH, Nienstaedt H (1953) Breeding timber chestnuts for blight resistance. Northeastern Forest Tree Improvement Conference Proceedings 1:70-75.

Griffin GJ, Hebard FV, Wendt RW, Elkins JR (1983) Survival of American chesntnut trees: evaluation of blight resistance and virulence in *Endothia parasitica.* Phytopathology 73:1084-1092.

Hebard FV (2005) Meadowview Notes 2004-2005. Journal of the American Chestnut Foundation 19(2):16-29.

Irwin JAG, Musial JM, Mackie JM, Basford KE (2003) Utility of cotyledon and detached leaf assays for assessing root reactions of lucerne to Phytophthora root rot caused by *Phytophthora medicaginis.* Australas. Plant Pathol. 32:263-268.

Kubisiak TL, Hebard FV, Nelson CD, Zhang J, Bernatzky R, Huang H, Anagnostakis SL, Doudrick RL (1997) Molecular mapping of resistance to blight in an interspecific cross in the genus *Castanea*. Phytopathology 87(7):751-759.

Kubisiak TL, Nelson CD, Staton ME, Zhebentyayeva T, Smith C, Olukolu BA, Fang G-C, Hebard FV, Anagnostakis S, Wheeler N, Sisco PH, Abbott AG, Sederoff RR (2013) A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*). Tree Genet. Genomes 9:557-571

Liebhard R, Koller B, Patocchi A, Kellerhals M, Pfammatter W, Jermini M, Gessler C (2003) Mapping quantitative field resistance against apple scab in a ‘Fiesta’ × ‘Discovery’ progeny. Phytopathology 93:493-501.

Newhouse AE, Spitzer JE, Maynard CA, Powell WA (2014) Chestnut leaf inoculation as a rapid predictor of blight susceptibility. Plant Dis. 98(1):4-9.

Parke JL, Roth ML, Choquette CJ (2005) Detached-leaf assays with *Phytophthora ramorum*: are they valid? Proceedings of the Sudden Oak Death Second Science Symposium.

Powell WA, Morley P, King M, Maynard CA (2007) Small stem chestnut blight resistance assay. Journal of the American Chestnut Foundation 21(2):34-38.

Tahi M, Kebe I, Eskes AB, Ouattara S, Sangare A, Mondeil F (2000) Rapid screening of cacao genotypes for field resistance to *Phytophthora palmivora* using leaves, twigs, and roots. Eur. J. Plant Pathol. 106:87-94.

Tedford EC, Miller TL, Nielsen MT (1990) A detached-leaf technique for detecting resistance to *Phytophthora parasitica* var. *nicotianae* in tobacco. Plant Dis. 74(4):313-316.

Table 2.1 Numbers of trees sampled in 2014 and 2015 at five sites with number of trees scored using the detached leaf assay and traditional stem inoculation.

Species/site	Leaf inoculation		Stem canker length		Stem rating	
	2014	2015	2014	2015	2014	2015
SIPAC-B3F2[a]	100	135	78	72	61	91
PWP-B3F2[b]	100	0	0	0	0	0
LF-B3[c]	5	39	0	34	0	0
LF-*Cm*[d]	5	9	0	5	0	0
MF-*Cd* [e]	5	5	0	0	0	0

[a] F2 progeny of open-pollinated third-backcross trees (B3F2) grown at Southern Indiana Purdue Ag Center (SIPAC);

[b] B3F2 trees grown at Potawatomi Wildlife Park (PWP); [c] Third-backcross (B3) trees grown at Lugar Farm (LF); [d] *Castanea mollissima*; [e] *Castanea dentata* at Martell Forest (MF).

Table 2.2 Mean leaf lesion lengths for tested pedigrees and sites.

Site/Species	2014	2015
PWP-B3F2[a]	35.14 a*	---
MF-*Cd*[b]	27.47 ab	32.9 a
LF-B3[c]	19.94 bc	32.17 a
SIPAC-B3F2[c]	19.87 bc	29.31 a
LF-*Cm*[d]	8.19 c	10.71 b

[a]Third-backcross F2 trees (B3F2) grown at Potawatomi Wildlife Park (PWP); [b]*Castanea dentata*; [c]Third-backcross (B3) trees grown at Lugar Farm (LF); [d]B3F2 trees grown at Southern Indiana Purdue Ag Center (SIPAC); [e]*Castanea mollissima*.

*Means followed by the same letters were not significantly different according to the Tukey's HSD test.

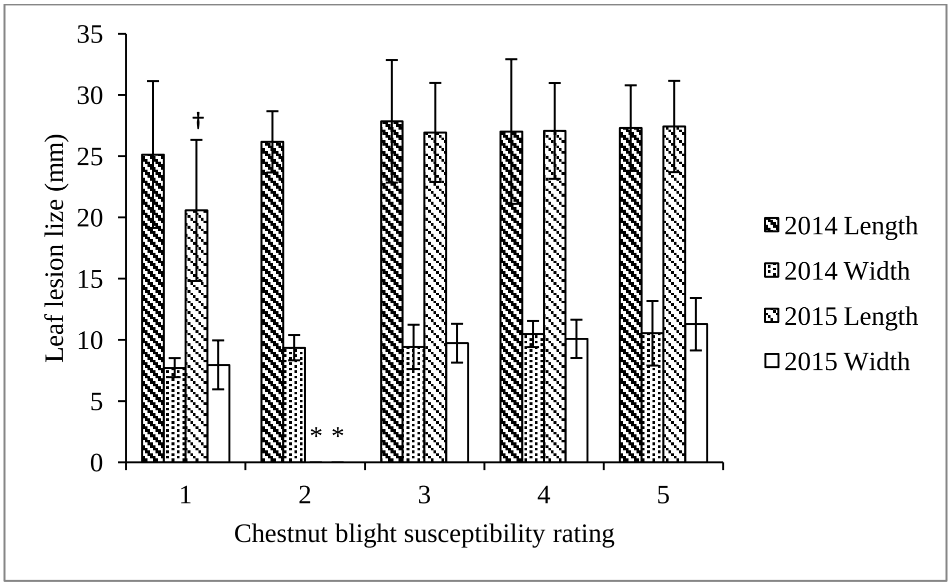
Table 2.3 Summary results of simple linear regressions of leaf and stem canker dimensions among B3F2 chestnuts at SIPAC

	2014			2015		
Model	b_1	p	r^2	b_1	p	r^2
SL[a]=LL[b]	0.12	0.023*	0.07	-0.02	0.003*	0.06
SW1[c]=LL	0.11	0.064	0.04	-0.01	0.41	0.01
SW2[d]=LL	1.15	0.518	0.01	-0.25	0.67	0.00
SL=LW[e]	0.26	0.202	0.02	0.29	0.026*	0.06
SW1=LW	0.05	0.032*	0.06	-0.01	0.428	0.01
SW2=LW	2.59	0.044*	0.06	0.24	0.457	0.01

[a]Length of stem lesion parallel to trunk; [b]Length of leaf lesion parallel to midvein; [c]Width of stem lesion perpendicular to trunk; [d]Width of stem lesion perpendicular to trunk, adjusted for diameter; [e]Width of leaf lesion perpendicular to midvein.

*p-value less than 0.05

Figure 2.1 Means of leaf lesion size among B3F2 chestnuts at SIPAC that received blight resistance ratings after stem inoculation. Error bars display standard deviation.



† Trees rated 1 in 2015 for stem susceptibility had a significantly different mean leaf lesion length versus trees rated 2, 3, 4, or 5 in 2015 as determined by a Tukey HSD test: there were no differences in means of lesion length or width in 2014 or 2015.

*Trees rated 2 in 2014 all received higher (more susceptible) ratings in 2015.

CHAPTER 3. GENETIC VARIANTS ASSOCIATED WITH CHESTNUT BLIGHT RESISTANCE AND SIGNATURES OF SELECTION NEAR BLIGHT RESISTANCE LOCI

Abstract

Understanding the genetic basis of resistance to forest diseases is important if developing disease-resistant planting stock, with the ultimate goal of ecological restoration, is a management goal for a disease-affected tree species. Using the Chinese chestnut reference genome for assembly of short reads, we identified resistance-associated polymorphisms in the genomes of 24 chestnuts with variable resistance to the necrotrophic canker-forming ascomycete *Cryphonectria parasitica.* Further analysis of genome regions containing blight-associated polymorphisms revealed predicted genes with plausible roles in blight resistance, including some that had >90% sequence identity with differentially-transcribed genes from canker-infected chestnuts in a previous study (Barakat et al. 2012) and some that map to previously identified blight QTL locations. Candidate genes for chestnut blight resistance include genes likely to be involved in pre-formed defenses and hormone signalling pathways, as well as receptor-like kinases and NBS-LRR-type proteins. Patterns of sequence variation were highly variable, with some loci showing clear evidence of low genetic diversity (strong selection) in the most resistant Chinese chestnut and others displaying high nucleotide diversity and heterozygosity in the most resistant trees. Evidence from our association analysis and chestnut transcriptomes (Barakat et al. 2012, Serrazina et al. 2015) indicates that an R-gene mediated resistance pathway may be activiated in susceptible American chestnut in response to blight that is not activated in resistant Chinese chestnut.

3. 1 Introduction

Introduced, virulent forest pathogens have caused enormous damage to forest ecosystems throughout the world. In the most extreme cases, such as the epidemics of chestnut blight and Dutch elm disease in North America, the affected tree species were largely eliminated from their former roles in the forest ecosystem, acquiring a new life history as a short-lived species (Anagnostakis 1987, 2001). In such cases, there is considerable interest in restoring the lost species by breeding enhanced disease resistance in survivors. In cases where the affected native tree has no heritable resistance to the introduced pathogen, hybrid breeding with a resistant congener is a way to incorporate resistance into the gene pool. Ideally, breeders should have an idea of how many loci control the disease resistance trait in their breeding population before attempting such a breeding program. The experiments necessary to determine this information in tree species are difficult to undertake, mainly because making crosses and evaluating phenotypes are less consistently successful and more time-consuming than in annual crops. A hybrid breeding program stands a greater chance of success the fewer genes are necessary to confer the desired level of resistance (e.g. Burnham et al. 1986).

Fortunately for tree breeders, there is considerable evidence that single genes, or small numbers of genes, are sufficient to confer resistance to pathogens in some plant-pathogen systems, a phenomenon known as gene-for-gene resistance. In gene-for-gene resistance, a gene that confers resistance in a plant corresponds directly to a gene that confers virulence in the pathogen (Agrios 2005). The virulence gene, or virulence factor "effector" in the pathogen is typically a protein that interacts with a cell-surface protein in the plant, so this form of disease resistance is also known as effector-triggered immunity (ETI). If the plant host possess an R gene allele that can recognize the effector, and the effector gene in the pathogen has an avirulent allele, a disease resistance response will be initiated and infection will not proceed. If the pathogen possesses an allele for the effector that the plant cannot recognize, or the plant lacks the R gene to recognize the effector, infection will proceed (susceptibility). The cell surface protein that actually interacts with the effector may or may not be the R gene. Gene-for-gene resistance was first demonstrated in the flax/flax rust system by H.H. Flor in the early 20th century, and it seems to be prevalent in plant-pathogen systems involving rusts. Since Flor's groundbreaking work, many R genes have been mapped, cloned, and sequenced. They are generally signal-transducing proteins with one or more of the following features: a transmembrane domain, an extra- or intra-cellular leucine-rich repeat (LRR) domain, an intracellular kinase, a nucleotide binding site, and a receptor domain. Nucleotide binding site – leucine-rich repeat (NBS-LRR) genes, a class of genes frequently implicated in gene-for-gene resistance, usually include a toll-interleukin receptor (TIR) or coiled-coil (CC) domain as molecular receptors. NBS-LRR proteins often do not directly interact with a corresponding avirulence gene in the pathogen, but rather are thought to act as guards that detect interactions with a different cell surface protein (Van Der Biezen and Jones 1998). In woody plants, NBS-LRR genes are thought to underlie a QTL for leaf-rust resistance in coffee trees (*Coffea* spp.) (Ribas et al. 2011). R genes involved in gene-for-gene resistance confer drastic differences in disease susceptibility phenotypes, so major-gene resistance, vertical resistance, and qualitative resistance are terms sometimes used to describe their phenotypic effect. Major genes for fusiform and white pine blister rust resistance are found in loblolly pine (*Pinus taeda*) and western white pine (*P. monticola*), respectively, indicating a gene-for-gene resistance system (Sniezko et al. 2014).

The resistance reaction that results from pathogen recognition by this type of R gene is characterized by the hyper-sensitive response (HR), a rapidly enacted programmed cell death reaction that, if successful, prevents the pathogen from colonizing tissue beyond its initial infection point (Morel and Dangl 1997). The evolutionary dynamics of gene-for-gene resistance have been characterized as a molecular “arms race” where rapidly evolving resistance genes in plants seek to counteract rapidly evolving avirulence genes in fungal pathogens (Boller and Yang He 2009). Evidence for the “arms race” is found in the large number of multiple-copy R gene clusters in plant genomes, the widely variable numbers of R genes found in different plant genomes (Wei et al. 2016), large numbers of alleles at R gene loci (Bakker et al. 2006), and the elevated nucleotide and amino acid diversity observed in the less-evolutionarily-constrained parts of R gene sequences (the LRR domain) (Rose et al. 2004, Thakur et al. 2013). Gene-for-gene resistance, and NBS-LRR genes underlying disease resistance QTL, are not limited to systems involving rust fungi. Diverse pathogens including apple scab (*Venturia inaequalis*) and late blight of potato (*Phytophthora infestans*) seem to interact with their plant hosts in a gene-for-gene manner (Soriano et al. 2009, Fry 2008). These pathogens are all biotrophs, which parasitize living plant tissue, or hemi-biotrophs, which begin by colonizing living plant tissue and transition to a necrotizing phase later on (Lee and Rose 2009). It seems gene-for-gene resistance is most effective when the pathogen’s lifestyle involves attempting ingress of living plant tissue without killing it: the programmed cell death of the HR is an effective strategy against pathogens that depend on living plant cells to grow (Mayer et al. 2001). Fully biotrophic pathogens must strike a balance between absorbing enough nutrients from the plant to complete their life cycle while remaining inconspicuous enough to avoid detection by the plant's immune system. Important invasive pathogens like white pine blister rust (*Cronartium ribicola*), root rots of many species caused by *Phytophthora cinammomi*, and sudden oak death (*Phytophthora ramorum*) (Hayden et al. 2014) all share biotrophic or hemibiotrophic lifestyles.

In addition to gene-for-gene or vertical resistance, many plants display resistance to pathogens that is inherited as a polygenic or quantitative trait. This phenomenon is also known as horizontal resistance, and it is often found in the same plant-pathogen systems as vertical disease resistance genes, including *Phytophthora* (Van Der Plank 1966, Nelson 1978). In horizontal resistance, a larger number of individual gene loci contribute to additive or incremental differences in disease resistance. The kind of complete resistance or immunity that comes with vertical resistance is rarely achieved by horizontal resistance; typically, disease symptoms are lessened rather than completely eliminated. The benefit of this is that an additive resistance allele at a given horizontal resistance locus in the plant does not impose intense selective pressure on the pathogen, as do R-genes that control vertical resistance in the gene-for-gene system. For this reason, plant breeders consider horizontal resistance to be more durable; incremental shifts in pathogen virulence over time will have less catastrophic effects on crop yields than a sudden shift to a new, virulent allele at a gene-for-gene locus (Parlevliet and Zadoks 1977). Horizontal resistance can be found in plant-biotroph systems, but it is more essential for breeding plant resistance to necrotrophic pathogens because vertical resistance to necrotrophic pathogens is rare (Poland et al. 2008).

Necrotrophs have adopted a wide range of strategies for their initial attack on the host. Some utilize appressoria or other structures to penetrate sound host cells, while many others exploit pre-existing wounds for access (Agrios 2005). In general, necrotrophs utilize a combination of phytotoxins, reactive oxygen species, and cell-wall degrading enzymes to kill host tissue (Laluk and Mengiste 2010). Some necrotrophs exploit mechanisms of the hypersensitive response, allowing host cells to kill themselves

(Shi et al. 2016). Necrotrophs can be divided into host specialists that are only virulent on one or a few host plant species (such as *Cochliobolus carbonum* and *Cryphonectria parasitica*), and host generalists that attack a broad range of plant hosts (*Botrytis cinerea* and *Sclerotinia sclerotiorum*). Host-specificity in some groups of necrotrophs depends on the production of host-specific toxins that are only effective against the host plant (Friesen et al. 2008). Corresponding to the diverse infection strategies of necrotrophic pathogens are diverse and complex mechanisms for responding to necrotroph infection in plants (Laluk and Mengiste 2010, Wang et al. 2014). Rather than depending on the recognition of effectors (ETI), resistance to necrotrophs often depends on the recognition of essential fungal molecules that are produced during infection (pathogen-associated molecular patterns or PAMPs) (Bent and Mackey 2007) or detoxification of pathogenic toxins (Poland et al. 2009). Recognition of necrotroph-produced molecules by the plant's pattern-recognition receptors can lead to PAMP-triggered immunity (PTI). Similarly, molecules associated with plant cellular damage, such as cell wall monomers (damage-associated molecular patterns or DAMPs), can trigger resistance responses. The classic HR response associated with ETI is not generally effective against necrotrophs; effector-triggered susceptibility has been observed in some necrotroph systems (Faris et al. 2010). The effector toxin victorin produced by *Cochliobolus victoriae* (a pathogen of oats) induces an HR response that leads to susceptibility (Lorang et al. 2012), and *Parastagonospora nodorum* exploits a receptor-kinase molecular pattern recognition pathway and the HR to cause disease in wheat (Shi et al. 2016). *Botrytis cinerea* can apparently exploit salicylic acid signaling pathways, associated with resistance to biotrophs, to cause a susceptible reaction (Rahman et al. 2012). Some necrotrophic pathogens show a gene-for-gene resistance response mediated by an NBS-LRR protein in *Arabidopsis* (Dobon et al. 2015). Effective resistance to necrotrophs involves the production of plant compounds that neutralize the phytotoxins or cell-wall degrading enzymes required for necrotroph pathogenesis. The signaling pathways that lead to resistance involve membrane-localized receptor-like kinases (RLKs), cellular mitogen-activated protein kinases (MAPKs), pathogenesis-related (PR) proteins, the hormone jasmonic acid (JA), and a variety of transcription factors. Since resistance in necrotrophs does not usually depend on the recognition of a highly variable, specific effector, elevated genetic diversity at gene loci involved in necrotroph resistance may not be as widespread as it is across conventional R gene loci. Resistance-associated genes in *Arabidopsis* that are not directly involved in pathogen recognition tend to undergo purifying selection and correspondingly have low nucleotide and amino acid diversity (Bakker et al. 2008).

Necrotrophic pathogens that attack fruit, leaves, and green tissue of herbaceous plants are known for causing rot diseases, which lead to the maceration, necrosis, and collapse of large areas of plant tissue (Agrios 2005). In woody plants, several economically and/or ecologically important canker diseases are caused by necrotrophic fungi, many of which cause stem canker diseases. The pathogen infects the outer, living layers of vascular tissue in wood stems, causing necrosis and an area of sunken, dead tissue. If the canker is not contained, it may cause death of branches, or the entire tree, by girdling the stem and cutting off water supply above the canker. Canker diseases caused by necrotrophs and hemibiotrophs include valsa canker of apple (*Valsa mali*) (Yin et al. 2016) and stone fruits (*Leucocytospora cincta*), a canker disease of Mediterranean oaks (*Biscogniauxia mediterranea*) (Moricca et al. 2016), pitch canker of pines (*Fusarium circinatum*) (Swett et al. 2016), and chestnut blight (*Cryphonectria parasitica*). Chestnut blight is an extremely virulent pathogen on American chestnut and nearly eliminated the species from forest ecosystems it once dominated, in concert with the introduced *Phytophthora cinammomi* in lower-elevation southern parts of its range (Anagnostakis 2001). While Chinese chestnut is variable in resistance to the pathogen (Qin et al. 1999), the most susceptible Chinese chestnuts are far more resistant than any American chestnut; i.e., a susceptible Chinese chestnut may lose large branches to blight cankers, but death of all above-ground tissue, as observed in American chestnut, is not generally seen.

The mechanisms of infection and host resistance in the chestnut/chestnut blight system are of interest because of the ongoing effort to develop blight-resistant American chestnut by introgressing resistance genes from Chinese chestnut; the likelihood of success in this endeavor depends on the nature of blight resistance in Chinese chestnut, and what exactly causes American chestnut to exhibit such a susceptible reaction. It is known that, whatever the host, chestnut blight infections typically start in bark wounds, frequently in the area around branch junctions (Metcalf 1912). Dissections of chestnut tissue undergoing chestnut blight canker development show death of host cells in advance of expanding *Cryphonectria* hyphae, indicative of a necrotrophic habit (Hebard et al. 1984). Initially, hyphae primarily colonize the phloem. The initial response to infection involves the formation of a lignified zone of dead cells around the fungal hyphae. Next, some hyphae grow into the lignified zone, while the tree attempts to encircle the hyphae with wound periderm. The fungus escapes encirclement by forming mycelial fans, which exert considerable pressure on plant tissue and lead to separation of phloem elements and deformation of the bark (Hebard et al. 1984). These mycelial fans do not penetrate wound periderm, but exploit gaps in the periderm while it forms to access tissue outside the initial lignified zone. Mycelial fans underlie the necrotic cankers that lead to the death of stems in susceptible trees; in susceptible trees, these fans penetrate to the level of vascular cambium. At the molecular level, mechanisms for infection and resistance are less well-understood. Oxalic acid is produced by *Cryphonectria parasitica* and functions as a phytotoxin as well as helping to break down cell walls (McCarroll and Thor 1978), but several other proposed toxins, including potential host-specific toxins, do not seem to kill chestnut tissue (McCarroll and Thor 1985a). Cell wall-degrading enzymes, namely polygalacturonase, appear to be important for breaking down host cells during the initial infection phase (McCarroll and Thor 1985b). Extracts of bark from American and Chinese chestnut both inhibit activity of *C. parasitica* polygalacturonase (Shain and Gao 1995), but Chinese chestnut bark extracts more strongly inhibit the enzyme. This inhibition is apparently not due to constitutively expressed tannins, which are somewhat more abundant in Chinese chestnut (Anagnostakis 1992).

The inheritance of chestnut blight resistance indicates a pattern of polygenic (horizontal) resistance (Jaynes 1974). Several genes are believed to contribute to the trait; this hypothesis has been borne out by QTL mapping experiments in interspecific hybrids that identified three major QTL for blight resistance on three different linkage groups (Kubisiak et al. 1997, 2013). Although no clear example of vertical resistance or near-immunity to blight seems to be present in Chinese chestnut, American (and to a lesser extent, European) chestnut displays a form of vertical susceptibility; members of the species are almost uniformly susceptible. While it is considered likely that more than three loci are involved in conferring chestnut blight resistance, it is also thought that only a few are necessary to raise the resistance of American chestnut to a level approaching that of Chinese chestnut (Kubisiak et al. 1997). mRNA sequencing analysis has shown that large shifts in transcription take place in tissue near cankers vs. healthy stems, and that transcriptional responses differ among American and Chinese chestnuts (Barakat et al. 2009, 2012). Genes related to phenylpropanoid biosynthesis, plant hormone signaling, and hypersensitive response were among the groups showing significant increases in transcription in infected stems of both species. Genes related to cell wall deposition, hydrolases, and oxidoreductases were over-represented in Chinese chestnut relative to American chestnut (Barakat et al. 2012). The individual genes underlying the difference in phenotypes, however, remain unknown.

To introgress blight resistance from Chinese chestnut into American chestnut, the American Chestnut Foundation and its collaborators backcrossed a pair of (*Castanea dentata* × *mollissima*) × *dentata* backcrossed chestnut hybrids (‘Clapper’ and ‘Graves’) to American chestnut for two generations to derive BC3 trees that, in theory, have genomes that are about 94% American chestnut in origin (Burnham et al. 1986, Hebard 2005). The BC1 resistance donors were selected from plantings established as part of an earlier chestnut resistance breeding program; backcrossing is thought necessary to recover the “timber-type” upright form and large stature of the American chestnut from hybrids, which tend to show the short, branchy habit of Chinese chestnut (Burnham et al. 1986). Since only resistant trees from each backcross generation are bred, the BC3 generation should have one American and one Chinese allele for blight resistance at each of the two or three loci that accounted for 75% of the variation in blight resistance observed by Kubisiak et al. (1997, 2013). When these trees are intercrossed, some proportion of the offspring will inherit two Chinese alleles at each resistant locus. These individuals should be true-breeding for blight resistance, i.e., their offspring will all be equally resistant to blight. Another generation of testing, of course, is necessary determine which of the best individuals are in fact true-breeding.

The success of the American chestnut restoration breeding effort depends on recovering nearly all of Chinese chestnut’s blight resistance in advanced backcross progeny. Since there is considerable variation in blight resistance among individual

Chinese chestnuts, choosing the best possible resistance donors would increase the likelihood of meeting the program’s goals. To this end, the American Chestnut Foundation added a number of highly resistant Chinese parents to its breeding program at Meadowview; incorporating more resistance donors remains a priority for ACF’s breeding program (Hebard 2006). Since the original resistance donors were BC1 trees, they contained at most one Chinese chestnut allele at each resistance locus. Many state chapter breeding programs use only ’Clapper’ or ’Graves’ as resistance donors. With a single BC1 resistance donor, true-breeding offspring of crosses between BC3 trees will be homozygous at resistance loci with two copies of a single allele from the Chinese grandparent of ’Clapper’ or ’Graves’. If both ‘Clapper’ and ‘Graves’ were included in a pedigree, most of the offspring would be heterozygous (C/G), but only two Chinese sources of resistance would be present. This is assuming that “Clapper” and “Graves” possess Chinese resistance alleles at the same 2 or 3 resistance loci, which may not be the case.

Whether or not reliance on a single version (allele) or a few versions of resistance genes is a liability for the restoration of American chestnut depends on the molecular basis of the differences in blight resistance observed among Chinese chestnuts. Since blight resistance has a strong effect on fitness, it is possible that resistance loci have evolved under purifying or negative selection. Given the horizontal nature of resistance to chestnut blight and the nature of resistance to necrotrophic pathogens in general, this seems possible; in fact, it is the most likely scenario if the genes involved are parts of a resistance pathway not directly involved in pathogen recognition. Only a few resistance donors would be needed, but selecting the most resistant Chinese chestnut parents available would still be essential.

The other possibility is that blight resistance loci have undergone balancing, or positive, selection. Positive selection would occur if unique alleles conferred an advantage against certain strains of the fungus; i.e., if they are molecular pattern recognition receptors (PRRs), such as transmembrane RLK or even NBS-LRR resistance genes. If high nucleotide and/or amino acid diversity is present in genome regions associated with blight resistance, incorporating a large number of Chinese parents would be essential to successfully developing blight-resistant chestnuts for restoration to eastern

U.S. forests, because the pathogen would be easily able to overcome one or two classical R genes in the restoration population.

The research questions addressed here are: 1) Is there evidence of balancing or purifying selection in the three genomic regions previously associated with blight resistance? 2) In, or near, which predicted genes are polymorphisms most strongly associated with blight resistance located? 3) What regions of the genome outside of the established resistance QTL show statistical associations with blight resistance, and what patterns of sequence diversity are found in those regions? 4) Do any of the predicted genes associated with differences in blight resistance have support from previously published transcriptome data?

To investigate these questions, we assembled whole-genome sequences of 24 individual chestnuts comprising highly resistant Chinese chestnuts, relatively susceptible Chinese chestnuts, highly susceptible American chestnuts, F1 hybrids of susceptible and hybrid species, and the BC1 'Clapper,' the resistance donor for many breeding populations in the eastern United States. Obtaining whole-genome sequences, while costly, allowed us to investigate patterns of association and sequence variation at a high level of detail. In particular, they allowed a high-resolution analysis of genes and genomic regions under selection (Lam et al. 2010, Slavov et al. 2012), even with a relatively small sample size (Guo et al. 2013).

3. 2 Materials and Methods

3.2.1 Plant Material

Trees were chosen primarily from the germplasm collection of the Empire Chestnut Company in Carrollton, OH (Table 1), including the highly resistant cultivar 'Nanking,' with two American chestnuts from a germplasm collection at Purdue University's Martell Forest (IN), and 'Clapper' and 'Mahogany' from the American Chestnut Foundation in Meadowview, VA. Initially, leaves were selected for DNA isolation, but dormant twigs were the source of most DNA sequenced because they resulted in better-quality samples. Resistance phenotypes for the trees chosen were based on long-term observations of performance in the field rather than artificial inoculations. With the exception of the two American chestnuts, all of the sampled trees have grown in chestnut orchards with a high level of natural exposure to *Cryphonectria parasitica* spores due to large amounts of fruiting *C. parasitica* on dead branches and stem cankers. Chinese chestnuts designated “susceptible” showed large (> 6” long) cankers on the trunk and branches, and death of large, prominent branches in the crown; hybrids designated “susceptible” showed dieback and resprouting. Trees were designated “resistant” if they showed no loss of major branches to blight cankers and no cankers > 6” in length.

3.2.2 DNA Isolation

DNA was extracted from twigs and leaves following the same protocol, which proved more effective for twigs. Plant tissue (one entire leaf or a three-inch section of first-year twig) was ground to a fine powder in liquid nitrogen using a mortar and pestle. The powder was placed in five mL of heated (50° C) CTAB extraction buffer and incubated four to eight hours at 50° C. Following incubation, one mL of 20 mg/mL proteinase K solution was added and samples were incubated for an additional 15 minutes. Five mL of 25:24:1 phenol:chloroform solution was added and samples were purified using a standard phenol:chloroform extraction (Doyle and Doyle 1987) followed by precipitation of DNA using 0.2 M sodium chloride and isopropanol. After pelletting and resuspending samples in TE buffer, contaminants were removed using Zymo Research OneStep PCR Inhibitor Removal kits (Zymo Research). Following purification, samples were quantified using a Nanodrop 8000 (ThermoFisher Scientific), and 2% agarose gel, then submitted to the Purdue Genomics Core Facility for sequencing.

3.2.3 DNA sequencing

Sequencing of 100 bp paired-end reads was carried out with an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA) at the Purdue Genomics Core Facility. In order to obtain >10x coverage of the ~800 Mb chestnut genome, two samples were sequenced per lane. Low-quality reads were filtered prior to assembly using Trimmomatic version 0.32 (Bolger et al. 2014).

3.2.4 Assembly of cbr QTL regions

The cbr1, cbr2, and cbr3 QTL scaffold sequences (Staton et al. 2014) were downloaded from http://www.hardwoodgenomics.org/chinese-chestnut-genome. cbr1-3 represent the three blight resistance QTL described in Kubisiak et al. (1997); each scaffold set includes individual genome scaffold sequences that align to the chestnut linkage map within or near the markers that define the blight resistance QTL. They were made publicly available as part of the Chinese chestnut reference genome sequencing project in 2012. Since there were several hundred relatively short scaffolds representing each QTL, scaffolds were concatenated with a 300 bp spacer of "N" or missing data between each scaffold to avoid reads bridging junction between sequences, which would result in many spurious polymorphisms due to assembly error around the junctions between individual scaffolds. Since the true order of the scaffolds was unknown, they were concatenated in descending order based on length (longest first). Scaffolds were concatenated separately, resulting in one reference sequence for cbr1, one for cbr2, and one for cbr3, which were analyzed separately. Short reads were assembled to the concatenated reference sequences using the Burrows-Wheeler aligner (bwa) (Li and Durbin 2009); alignments were sorted and duplicate reads removed using Picard Tools, and realignment around indels and calling of SNP and indel polymorphisms was carried out using the Genome Analysis ToolKit (GATK) (McKenna et al. 2010) best practices pipeline minus the variant quality score recalibration step (DePristo et al. 2011, Van der Auwera et al. 2013). SNPs were filtered for depth (greater than 5, less than 100) and quality (read quality averaged over samples > 40) using VCFtools (Danecek et al. 2009). The purpose of the depth filter was to ignore repetitive sequences with depth inflated due to spuriously assembled low-complexity reads.

3.2.5 Assembly of chloroplast and whole genome

Chloroplasts were assembled by assembling short reads to the complete Chinese chestnut chloroplast reference sequence (Jansen et al. 2011). The 1.0 version of the Linkage Group A (LGA) pseudochromosome assembly and beta versions of the LGB-LGL assemblies (12 total) (Staton et al. 2014) were obtained from Dr. John Carlson of Penn State University. Reads were assembled to the whole draft genome using bwa;

duplicate filtering, polymorphism calling, and quality filtering of polymorphisms were carried out using the same protocols as for the cbr-QTL sequences.

3.2.6 Gene prediction and filtering

De novo gene prediction was carried out separately for the cbr-QTL reference sequences and for the whole genome using AUGUSTUS (Stanke et al. 2006) with *Arabidopsis thaliana* as the training protein set and default settings. To assign a putative function to predicted genes, the predicted gene file (.gff) was converted to fasta (.fa) format and aligned to the UniProt protein database using the blastp function of the DIAMOND sequence aligner (Buchfink et al. 2015) with default settings. The top hit was used to assign a putative function of the gene; in most cases, the functional annotation was based on alignment to a protein of known function in *Arabidopsis thaliana*. To provide a measure of validation to this predicted gene set, publicly available cDNA contig files for American chestnut, Chinese chestnut, European chestnut, and Japanese chestnut were downloaded from http://www.hardwoodgenomics.org/transcriptomes. These were each aligned using the blastx function of DIAMOND, with default settings, to a database generated using the predicted Chinese chestnut protein set output by AUGUSTUS. Transcripts were matched to the protein that provided the top hit from the predicted protein set; a predicted protein was only counted as having transcript support if it was the best alignment for at least one cDNA contig. This process was carried out using a custom Perl script. The list of alignments was also searched for cDNA contigs that were designated differentially expressed in Barakat et al. (2012). AUGUSTUS output was also converted to .bed format and used to filter .vcf files (VCFtools) for those polymorphisms that occurred in predicted gene sequences and in predicted exons. Genes predicted in the LGA pseudochromosome sequence using MAKER (Cantarel et al. 2008) were downloaded from the Hardwood Genomics website to provide additional information on the potential structure of genes predicted by AUGUSTUS.

3.2.7 Phylogenetic Tree Construction

Phylogenies were constructed for chloroplast and nuclear SNPs using the maximum-likelihood method in SNPhylo and visualized using SNPhylo (Lee et al. 2014) and the PhyloDendron online tree viewer (http://iubio.bio.indiana.edu/treeapp/).

3.2.8 Association Tests and Statistics

Association tests were conducted using the –perm function of the Plink software 1.07 and 1.09 (Chang et al. 2015), which uses an adaptive Monte Carlo permutation test to assess statistical significance. Association tests were performed separately for the three cbr-QTL sequences and for the twelve pseudochromosomes individually. For the association test, resistance was modeled as a qualitative (case/control) trait: 0 = resistant, 1 = susceptible.

Tajima’s D, pi, and heterozygosity were calculated using VCFtools. A custom Perl script was used to calculate statistics for large numbers of predicted gene sequences individually, and to analyze the potential amino acid changes SNPs and indels would cause. Candidate genes were selected based on the presence of associated SNPs and indels, their potential effects on protein products, plausible roles in blight resistance, and evidence of expression.

3.3 Results

3.3.1 Sequencing and assembly

The target 10x depth was attained for all but 2 individuals (Table 2), and these were each above 6x depth. Between 5-10% of reads were discarded by Trimmomatic due to low read quality prior to assembly. When SNPs were called on the whole-genome assemblies, a reasonable Ts/Tv ratio of about 2.6 was obtained for assemblies on all linkage groups (Table 3). After a quality (total quality > 1000) and depth (maximum average depth per individual < 45) was applied, 18,360,448 polymorphic sites remained across the genome.

3.3.2 Phylogenetic analysis

Tree construction from chloroplasts revealed a *Cc* chloroplast in some Korean-derived *Cm* material (individuals labeled NC; Figure 1), and a *Cd* chloroplast in 'Schmucki' and one other putatively *Cm* tree. The nuclear genomes of these individuals, however, indicated that they had been backcrossed to *Cm* for at least one generation as their position on a nuclear SNP tree indicated similarity to other Chinese chestnut (Figure 2). 'Clapper' had a unique *Cm* chloroplast relative to the other trees sampled, which all shared a single chloroplast haplotype.

3.3.3 Gene prediction

AUGUSTUS predicted 86,571 genes across the entire genome, which is two to three times higher than the typical eukaryotic gene count. When predicted protein sequences were aligned to the UniProt/SwissProt database of curated proteins, 37,458 were aligned to proteins from the database with e-value less than the default Diamond e-value cutoff of 0.001. This was still higher than expected, possibly because of the inclusion of a large number of predicted proteins that aligned to transposable element proteins and were unlikely to represent functional gene loci. AUGUSTUS also predicted pairs of genes on LGA in several places where MAKER predicted a single coding sequence. Filtering of polymorphism .vcf files for predicted genes was performed using the full set of predicted genes (including those without alignments); this led to a ~40% reduction in the number of SNPs and indels (Table 4). When polymorphisms were filtered to include only those within predicted exons of predicted genes, their number was reduced to about 25% of the original number of SNPs and indels from the whole-genome assembly and the Ts/Tv ratio increased to >3 (Table 5).

3.3.4 Association tests

A large number of associated SNPs and indels were found, even with stringent permutation-based P-value cutoffs (Table 3, Table 4, Table 5), but they tended to be clustered in short segments along each linkage group (Figures 3-15). Using a cutoff of 100 polymorphisms with a permutation p-value <0.005 in a 5000-polymorphism bin, five regions were identified on LGA, one each on LGB and LGC, two on LGD, one on LGE, three each on LGF and LGG, none on LGH and LGI, three on LGJ, two on LGK, and four on LGL. After examining the distribution of associated polymorphisms around predicted genes independent of the total number of associated polymorphisms in a given region (i.e., to find individual genes potentially associated with blight resistance), four loci were added for investigation on LGB and one on LGG, while several loci were not investigated further on LGF, LGJ and LGK because the locations of associated SNPs and indels relative to predicted genes did not have any clear biological interpretation. Predicted genes in the selected regions aligned to a variety of disease and stress-response proteins from *Arabidopsis* (Table 6), and to similar predicted proteins in the genomes of other woody plants, in particular Persian walnut (*Juglans regia*) (Table 7). Many statistically-associated polymorphisms occurred within retroelement-associated predicted genes on LGA (328 associated with $p < 0.001$) and LGL (292) in particular. A relatively small number of associated polymorphisms were predicted to confer amino acid changes (Table 8).

3.3.5 Alignment of transcript data to predicted proteins

Predicted proteins within blight resistance-associated genome regions were considered to have transcript support if they were the top hit of at least one chestnut transcript; using this method, 49 of 79 had transcript support from Chinese chestnut (Table 9), whereas 40 of 79 had support from American chestnut (Table 10). Of these, 13 (6 in American chestnut and 7 in Chinese chestnut) were the best alignment for at least one transcript that was differentially expressed in canker vs. healthy stem tissue in Barakat et al. (2012). Twenty-one of the predicted genes were also associated with transcripts from Japanese chestnut and 19 with European chestnut; 6 were associated with differentially expressed transcripts from *Phytophthora*-infected root tissue of European and Japanese chestnut (Serrazina et al. 2015).

3.3.6 Patterns of nucleotide divergence and heterozygosity in regions associated with blight resistance

When only the cbr scaffold sequences were used for assembly and SNP calling, the average Tajima’s D value over the entire set of scaffolds was somewhat higher in resistant Chinese chestnuts than in susceptible Chinese chestnuts or American chestnut and 'Paragon.' This pattern was observed at a number of the putative resistance loci elsewhere in the genome. While only seven of 79 genes selected in Table 12 show a negative Tajima's D value in highly resistant Chinese chestnuts, 35 and 36 had negative Tajima's D in susceptible Chinese chestnuts and highly susceptible species, respectively. Seventeen genes had Tajima's D values greater than 2 in resistant Chinese chestnuts, indicating a history of diversifying selection, while only six in susceptible Chinese chestnut had a value of D that high and none in highly susceptible species (Table 11). Conversely, using π as a measure of nucleotide diversity, the number of genes with higher diversity in resistant Chinese chestnuts was approximately equal (33) to the number with higher diversity in susceptible Chinese chestnuts (31). Forty-nine of 79 genes had higher heterozygosity in resistant Chinese chestnut than in susceptible trees. Heterozygosity was generally higher in *Cm* × 'Paragon' hybrids than in either species. Average F_{ST} among species was 0.49, with a minimum of 0.016 and a maximum of 0.964. F_{ST} was negatively correlated with heterozygosity in the most resistant trees (R^2=0.37); i.e., for those genes with the most strongly divergent genotypes between species, the most resistant Chinese chestnuts had low heterozygosity (Table 12).

3.3.7 Analysis of putative disease resistance loci by functional category

Predicted resistance genes in different functional categories varied in average Tajima's D and differentiation among species, as measured by heterozygosity in interspecific hybrids, based on their predicted functional categories (Table 13). NBS-LRR and lectin receptor kinases showed more evidence of diversifying selection, while DETOXIFICATION efflux transporters and cytochrome P450 genes showed more evidence of purifying selection and strong differentiation among species.

3.4 Discussion

3.4.1 Chloroplast and nuclear marker phylogenetic analysis

The results from the chloroplast analysis (Figure 1), which indicated that all of the Chinese chestnuts sampled had an identical chloroplast, supports the idea that most Chinese chestnut germplasm in the United States is derived from a limited gene pool.

The haplotype most frequently observed in this sample was also most frequently observed in populations of wild chestnuts from southern China (Chapter 5), which indicates that southern China is the most likely origin for most of the Chinese chestnut germplasm in North America. ‘Clapper’ had the only unique *C. mollissima* chloroplast in this sample; this haplotype is present in some southern Chinese populations but occurs at highest frequency in northern Chinese orchard material (Chapter 5).

A tree constructed using SNPs from the nuclear genome (Figure 2) shows ‘Clapper’ clustered with American chestnuts; as a BC1 tree, ‘Clapper’ is expected to have a genome that is 75% *Castanea dentata*. When a tree was constructed using SNPs from the cbr1 blight resistance QTL scaffolds, ‘Clapper’ clustered with American chestnuts; when trees were constructed for the other two QTL scaffold sequence, ‘Clapper’ clustered with ‘Paragon’ × *C. mollissima* hybrids, indicating that ‘Clapper’ has a hybrid genotype at cbr2 and cbr3, but two American chestnut chromosome segments across cbr1. ‘Clapper’ may not be an ideal blight resistance donor if the absence of a *C. mollissima* allele at cbr1 cannot be compensated for by other blight resistance loci. Several resistant and susceptible chestnuts with some Korean background (NC1, NC2, NC4, and NC6), which had a *Castanea crenata* chloroplast, also have distinct nuclear genomes (Figure 2), clustering together on a branch of the tree. Chinese chestnuts with American chestnut admixture as indicated by chloroplast data, however, clustered near other Chinese chestnuts, which indicates that backcrossing to Chinese chestnut has removed most of the American chestnut nuclear genome from these trees, one of which (‘Schmucki’) is exceptionally resistant to chestnut blight.

3.4.2 Association analysis

A large number of SNPs with statistically significant ($p< 0.005$) associations with blight resistance were observed, even when only SNP loci within predicted genes were considered. The large number is most likely due to the fact that six of the susceptible trees inherited a large portion of their genome from susceptible parent species (two American chestnuts, two ‘Paragon’ × *Cm* hybrids, ‘Paragon,’ and ‘Clapper,’ while only two (‘Schmucki,’ one ‘Paragon’ hybrid) of the resistant trees did. Some of the trees in the sample (‘Paragon’ and its offspring) had known family relationships. While most of the *Cm* individuals are believed to be unrelated, some of them are derived from a limited North American orchard gene pool that may have led to larger-than-expected haplotype blocks existing in the SNP dataset.

3.4.3 Patterns of sequence variation across regions associated with chestnut blight resistance

The nature of sequence variation at loci associated with blight resistance is important because it indicates whether or not breeding for homozygosity at blight resistance loci is a liability for a resistance breeding program. If sequence diversity across a given gene or locus (Tajima's D, π, heterozygosity) is lowest in resistant *Cm* and hybrids and higher in susceptible trees, the locus is most likely under purifying selection, and breeding for homozygosity would not be an issue. The final products of the breeding program would reflect the genetic condition of the most resistant trees in nature. On the other hand, if nucleotide diversity tends to be higher in blight-resistant trees and lower in susceptible trees, there may be an advantage to a larger number of resistance donors and higher genetic diversity at blight-resistance loci. Breeding for homozygosity at blight resistance loci would nullify this advantage. We considered measures of nucleotide diversity for a set of predicted genes that, based on statistical SNP associations with blight resistance, functional annotations, and transcript evidence, represent potential candidate genes for blight resistance in chestnut. Tajima's D statistic showed a signal of diversifying selection at many of the selected blight resistance candidate loci in the most resistant Chinese chestnut, but often not in less-resistant Chinese chestnut and almost never in susceptible chestnut species. Individual genes in LGA.a, LGA.c, LGB.c, LGC.a, LGE.a, LGK.a, LGL.b and LGL.d in particular showed this pattern, including LGA.d.2 and LGA.d.3, predicted receptor-like kinases, and LGB.e.5, an F-box protein. LGK.a.1, differentially expressed in Chinese chestnut (Barakat et al. 2012), and LGB.d.4, LGD.a.1, and LGG.d.3, differentially expressed in American chestnut, showed higher Tajima's D and π values in resistant Chinese chestnuts. Conversely, many loci showed no difference in Tajima's D statistic across groups, or was even lower in resistant Chinese chestnuts. This pattern may reflect purifying selection. Loci showing this pattern included LGA.c.1, LGA.e.1, LGA.e.2, LGG.b.1, and LGG.c.2, which all showed differential gene expression in cankers of Chinese chestnut (Barakat et al. 2012).

Several genes with elevated Tajima's D and/or π in resistant Chinese chestnut, relative to more susceptible trees, were pattern-recognition receptors, although the putative NBS-LRR gene LGL.c.2, which was differentially expressed in American chestnut cankers (Barakat et al. 2012) and *Phytophthora*-infected roots of European and Japanese chestnut (Serrazina et al. 2015), had a signature of neutral selection in Chinese chestnut based on Tajima's D. Other predicted genes with signatures of diversifying selection in resistant Chinese chestnuts were similar to known F-box protein-ubiquination proteins, which generally have a role in protein degradation (Ho et al. 2006). If these proteins are involved in degrading disease-resistance proteins to modulate the resistance response, it may be that resistant Chinese chestnuts have developed diverse alleles at these loci because the proteins interact with diverse pattern-recognition proteins that detect pathogen attack.

Genes with the highest number of associated SNPs were most often found in resistant Chinese chestnuts that had a unique allele found in neither more susceptible Chinese chestnut nor susceptible species. Resistant Chinese chestnuts typically were heterozygous for these alleles, which may indicate they cause reduced fitness in homozygotes. The presence of rare alleles tended to drive higher Tajima's D values in resistant Chinese chestnuts where they occurred. Genes with unique alleles probably conferred marginally improved (quantitative) resistance within Chinese chestnut, but may not affect the large (qualitative) difference in resistance between susceptible Chinese chestnuts and American and European chestnuts. These genes are identifiable by high heterozygosity in highly resistant *Cm* (het-CmR category of Table 13) relative to susceptible species and relatively susceptible *Cm*. Other loci with smaller numbers of statistically associated SNPs were fixed at one allele in resistant and susceptible Chinese chestnuts and a different allele in the susceptible species. Loci with this pattern of allelic variation, in particular on LGG, site of the original cbr3 blight resistance QTL, are more likely to confer the large difference in resistance between American and susceptible Chinese chestnuts. These genes can be identified by low values of Tajima's D in resistant Chinese chestnuts (Table 12), extremely high F_{ST} values among species (Table

13) and elevated heterozygosity in the ‘Paragon’ × Chinese chestnut hybrid category (Table 13).

The implications of these results for chestnut breeding are mixed. There appears to be elevated heterozygosity and nucleotide diversity in resistant Chinese chestnuts at a number of loci where polymomrphisms were statistically associated with blight resistance. However, genes with high interspecific F_{ST}, low genetic diversity in resistant Chinese chestnuts, and proximity to previously identified blight resistance QTL in pseudochromosomes corresponding to linkage groups B and G were also observed. Predicted genes with these characteristics may represent the most likely candidates for the large differences in resistance among chestnut species. A breeding strategy that created a genetic bottleneck for resistance alleles (i.e. one or two resistance donor alleles in a breeding population) would not be problematic if these loci are responsible for most of the interspecific difference in blight resistance. Conversely, the relatively large number of predicted disease resistance loci where resistant Chinese chestnuts show evidence of elevated allelic diversity would be disadvantaged in such a breeding strategy. Even if loci with higher genetic diversity in the most resistant Chinese chestnuts confer, on average, a marginal increase in blight resistance, maximizing blight resistance is crucial for the successful restoration of American chestnut on a landscape that hosts a large and genetically diverse population of *Cryphonectria parasitica*. As the number of target genes increasing, introgression via backcrossing becomes more difficult because transgressive segregants for all targeted genes will become less and less frequently observed. Focusing on incorporating as many Chinese chestnut resistance gene alleles as possible in individual lines and then intercrossing the best individuals from backcrossed lines could be a way for breeders to generate a restoration chestnut population with strong *Cm* allelic diversity at a larger number of resistance loci at the population level.

3.4.4 The molecular basis of blight resistance inferred from association analysis, predicted gene annotation and transcriptomic data

Plant disease resistance responses involve dramatic departures from the normal function of plant cells, and many individual proteins contribute to steps in disease response (Jones and Dangl 2006). The transcriptional reprogramming of healthy versus blight-infected chestnut stems (Barakat et al. 2009, 2012) is direct evidence of the complex molecular basis of chestnut blight resistance and susceptibility in chestnut. Despite the large number of genes involved in a tree's reaction to chestnut blight, only two or three are apparently necessary to confer blight resistance in interspecific hybrids of American and Chinese chestnut (Kubisiak et al. 1997). These genes might be involved in pre-formed defenses, or proteins that control key steps in transcriptional reprogramming during the response to chestnut blight. Of the SNP associations depicted in Figures 3 to 15 and associated candidate genes depicted in Tables 6 to 13, the best candidates fell into several categories of molecular function that, together, provide some evidence of the molecular basis of chestnut blight resistance, and what causes the drastic susceptibility of American chestnut (Table 16).

3.4.5 Pattern-recognition receptors

Pattern-recognition receptors (PRRs) are a large and diverse group of proteins that mediate defense reactions by detecting pathogens and triggering responses within the plant cell (Jones and Dangl 2006). In chestnut, we identified ten candidate genes that are similar to known PRRs in public protein databases. Genes that fall into this category occur at the loci LGA.d, a cluster of receptor-like kinases similar to wheat rust resistance loci, LGB.a, a lectin-receptor kinase and an NBS-LRR gene, and LGL.c, a large cluster of NBS-LRR genes. Possibly, LGA.e, which contains a large transmembrane protein, falls into this category as well, but the predicted genes underlying this locus are not homologous to known plant PRR or disease resistance proteins. PRRs show evidence of transcription in Chinese and American chestnut, but differential expression (upregulation) only in infected stems of American chestnut. The LGL.c locus, which contained a cluster of PRR-like predicted genes, also showed a signal of up-regulation in response to *Phytophthora* infection in European and Japanese chestnut (Serrazina et al. 2015). The runaway necrotic canker formation in American chestnut, accompanied by peroxidase activity and other hallmarks of programmed cell death and the HR (Barakat et al. 2012), indicates that an over-active HR may lead to American chestnut's susceptibility. Since none of these PRR loci appear among the original cbr QTL, it appears that Chinese chestnut genes other than the actual pattern-recognition receptors are able to "rescue"

American chestnut by attenuating the disease response with regulatory genes further downstream in the pathogen response. The most resistant Chinese chestnuts showed elevated Tajima's D at the LGA.d locus, which contained a small cluster of LRK10-like receptor kinases, so it may also be that the most resistant Chinese chestnuts have unique alleles at the loci that attenuate the HR more effectively than susceptible Chinese chestnuts.

3.4.6 Auxin , abscisic acid, ethylene, and jasmonic acid signalling

The role of plant hormones in disease resistance is well-documented (Denancé et al. 2013). Salicylic acid (SA), jasmonic acid (JA), and ethylene are the most important hormones in many *Arabidopsis* disease resistance reactions (Clarke et al. 2000); in general, SA pathways regulate responses to biotrophs, and JA/ethylene regulate responses to necrotrophs (Denancé et al. 2013). Other plant hormones, including auxin and abscisic acid (ABA) are primarily involved in regulating plant growth, but can be involved in disease resistance because of interactions with the SA and JA pathways (Denancé et al. 2013). Auxin signalling and metabolism was associated with predicted genes found at LGB.b and LGD.a, and also possibly LGB.c. Abscisic acid was associated with LGG.b and LGB.e, and also possibly LGA.b and LGB.c. Ethylene was associated with LGB.e and several ethylene-responsive transcription factors at LGG.d. Jasmonic acid, often associated with resistance pathways to necrotrophic pathogens, was only associated with LGB.b. Predicted genes that may be involved in hormone signaling include several on LGB and LGG that could underlie two of the original interspecific blight resistance QTL, cbr1 and cbr3. Of the seven predicted genes that had evidence for differential transcription in Chinese chestnut infected stem tissue, one was involved in auxin signalling and two were involved in ABA signalling. Among the predicted genes that had evidence of differential transcription in American chestnut infected stem tissue (Barakat et al. 2012), two different genes had a role in auxin or ABA signalling, and one was an ethylene-responsive transcription factor. Differences in auxin and ABA signalling pathways appear to influence the differential success of American and Chinese chestnut in responding to chestnut blight attack. ABA has been implicated in susceptibility to some pathogens, but resistance to others, due to its role in enhancement of callose deposition (Mauch-Mani and Mauch 2005). ABA can also influence the behavior of MAPK (mitogen-associated protein kinase) kinase signalling pathways that are crucial for disease resistance (Danquah et al. 2014).

3.4.7 Downstream regulation of defensive response

Beyond the initial detection of the pathogen, a number of genes in conserved disease response pathways transmit signals from cell surface proteins to enzymes that actually execute the disease response; generally,(MAPK) modules are crucial to transmit disease response signals (Meng and Zhang 2013). Two loci (LGG.d and LGL.b) contained predicted genes similar to the PBL27 serine-threonine protein kinase of *Arabidopsis*, which is involved in a MAPK-kinase cascade (Yamada et al. 2016). Since these genes and pathways are highly conserved, they are also good candidates for the difference in blight resistance among resistant and susceptible chestnut species.

3.4.8 Defense against fungal weapons

To defend against necrotrophic pathogens, plants may develop enzymes that degrade toxins produced by the pathogen (Mengiste 2012). Several of the predicted genes at associated loci appear to be involved in defending against specific elements of pathogen attack, including a pectinesterase inhibitor (on LGD), the BODYGUARD-like gene (LGA.b) that is involved in cell wall modification, and several efflux proteins, including a NIP51-like gene on LGA and the DETOXIFICATION-like gene cluster (LGB.c) that may include the causative gene underlying the cbr1 QTL. Few of the predicted genes with potential roles in defense against pathogen attack show differential expression in American or Chinese chestnut, but several, including the pectinesterase inhibitor, are expressed in Chinese but not American chestnut. In general, the genes with this potential molecular role are conserved within species and differentiated among species.

3.4.9 Role of transposable elements

The open reading frames associated with POL polyproteins of many transposable elements are interpreted as protein-coding loci by gene-prediction programs like

AUGUSTUS. Seven predicted retroelement-like genes were nearest-neighbors to predicted genes listed below as potential blight resistance candidates (LGA.b.1, LGA.d.2, LGA.d.3, LGB.a.1, LGB.e.2, LGD.b.2) and contained large numbers of statistically associated SNPs. Considering all of the putative blight resistance loci, a total of 58 predicted retroelement-like proteins were found within regions with the highest concentrations of blight-associated SNPs; these predicted genes encompassed 5947 polymorphisms statistically associated with blight resistance ($p < 0.005$) within their predicted exon sequences. It could be that these repetitive sequences are simply linked to polymorphisms in nearby genes or promoter-binding regions, or even that the repetitive nature of transposable elements led to an excess of spuriously assembled reads in these regions, inflating levels of polymorphism. It is also possible that these transposable elements are important for the differences in blight resistance phenotypes. The presence of transposable elements (TEs) can affect transcription in several ways: indirectly, via methylation of regions with transposable element insertions leading to differential transcription (Cui et al. 2014), or directly, causing up-regulation of a gene by modifying promoter or enhancer regions upstream (Negi et al. 2016). Transposable elements have been observed to re-activate genes by acting as promoters when inserted upstream (Hayashi et al. 2008). It is possible that differences in transposable element location, type, methylation state or activity cause some genes to be transcribed in American chestnut but not in Chinese chestnut or vice versa, leading to differences in blight resistance. At many of the blight resistance-associated loci and candidate genes identified, the majority of associated polymorphisms were either in non-coding portions of genes, in transposable element-related genes, or outside of gene sequences. Therefore, it appears likely that any difference in phenotype conferred by some candidate genes is due to differences in promoter or enhancer sequences, intron splice sites, or other non-coding elements rather than differences in amino acid sequence. In some cases, associated SNPs were located in proximal (<5000 bp) upstream control regions (Table 15) of candidate genes.

3.4.10 Detailed discussion of putative blight resistance loci

The functional annotations of genes potentially associated with blight resistance were mostly based on the annotation of the *Arabidopsis* genome. Several plausible loci and candidate genes are described below in detail.

The most promising candidates at the locus LGA.a are neighboring ALF4-like genes (predicted by AUGUSTUS) or one longer gene (predicted by MAKER). The predicted genes at this locus show moderately strong homology (44-48% amino acid identity) with the *Arabidopsis* ALF4 protein (Table 6) and stronger homology (64-68% amino acid identity) with predicted proteins from woody plants like *Juglans regia* (Table 7), and are supported by transcripts from all chestnut species surveyed except *Castanea sativa,* with 100% identities in *C. mollissima* and *C. crenata.* Named for its role in lateral root formation, ALF4 is also expressed in stems of *Arabidopsis thaliana.* It is not directly involved in auxin signalling (DiDonato et al. 2003). Because it is expressed in both roots and stems and causes distinct mutant phenotypes in each, ALF4 is believed to have some role in maintaing the ability of non-meristematic tissues, like the root pericycle, to undergo mitosis. Interacting proteins (Braun et al. 2011) for *Arabidopsis* ALF4 include TIFY8, HUB1 ubiquitin-protein ligase, UBQ3 polyubiquitin, and HSP23.6-MITO heat-shock protein. Several blight-associated SNPs that are predicted to cause amino acid changes occur within two predicted exons. One involves a serine-alanine substitution that could potentially influence the function of the protein; this SNP has a reference allele fixed in Chinese chestnuts and an alternate allele fixed in American chestnut. Two adjacent non-synonomous SNPs in the same exon have a distinct genotype in the most resistant *Cm*, while susceptible *Cm* share the reference allele with *Cd*. This locus also contains a gene (LGA.a.3) similar to TR120 in *Arabidopsis*, which encodes a subunit of the trafficking protein particle complex. TR120 is involved in post-Golgi protein trafficking and is crucial for normal development because it is required for transport of PIN2, an auxin efflux carrier, to the plasma membrane (Qi et al. 2011). LGA.a.3 (TR120-like) contains several nonsynonymous SNPs that are statistically associated with blight resistance, and is supported by transcripts from all four chestnut species, so it is fairly highly conserved. Conservation is also shown by its 80% amino acid sequence identity to the closest *Arabidopsis* homolog. Its putative role in blight resistance is based on its potential role in the auxin signaling pathway.

Most of the associated SNPs at the locus LGA.b occurred within the confines of a predicted transposon-related protein (AUGUSTUS) upstream of a BODYGUARD 2 (BDG2)-like lysophospholipase gene. The sequence and structure of this gene predicted by Maker and AUGUSTUS were similar. The BDG2 gene only contained one nonsynonomous SNP within its predicted exons, and this polymorphism did not have any statistical association with blight susceptibility, although several synonomous SNPs within the gene sequence did. It seems that polymorphisms at this locus most likely affect 5' regulatory regions of the BDG2 gene rather than its coding sequence. The BDG2 gene is supported by *Cd* and *Cm* transcripts, but not by *Cc* or *Cs*, which indicates that it may be involved in a specific response to blight rather than a general disease response. There are a number of BDG-type genes in *Arabidopsis*, and all are involved in modification to the cell wall and cuticle (Kurdyukov et al. 2006), so they could be involved in constitutively expressed defenses (a thicker cuticle) or cell-wall remodeling in reaction to pathogen attack. In particular, the expression of BDG genes in suberized regions (Jakobson et al. 2016) points to a potential role in canker containment as lignfied cells are modified to become a stronger suberized barrier to the blight fungus. BDG proteins may have a specific role in resistance to necrotrophic pathogens: in one study (Chassot et al. 2007), altering the BDG1 gene in *Arabidopsis* conferred resistance to *Botrytis cinerea*. The poorly formed cuticle in the BDG mutants apparently stimulated an effective defense response similar to that elicited by cuticle breakdown during pathogen attack. BDG proteins are induced by osmotic stress and abscisic acid (ABA) (Wang et al. 2011), and interact with other lipid-processing enzymes such as SPT1 and LCB2 serine palmitoyltransferases according to the STRING protein interaction database.

Few of the predicted genes at the LGA.c locus had convincing transcript support, but a probable aquaporin with homology to nodulin-26 intrinsic protein (NIP5-like) pore proteins in *Arabidopsis* did, and it showed differential (reduced) expression in healthy stem vs. canker tissues in *Cm* (Barakat et al. 2012). Of the predicted genes in blight-associated regions that aligned to differentially expressed transcripts from American and Chinese chestnut blight cankers, this was the only one that showed reduced expression in cankered stems. Of the 10 polymorphisms in predicted exons of this predicted gene, eight were non-synonymous, and all but one showed a pattern of fixation at a reference allele in *Cm* and fixation at an alternate allele in *Cd*. One of these nonsynonymous SNPs was predicted to cause a premature stop codon, and several others conferred major amino acid changes. Transcripts from all four surveyed chestnut species aligned to this protein with high similarity (Table 9, Table 10). The *Arabidopsis* protein with the strongest homology to the predicted chestnut protein, NIP5-1 (80% identity), is a boric acid transporter that is expressed in above-and below-ground tissues (Quigley et al. 2002) although some similar proteins are involved in the transport of highly toxic arsenite (Kamiya et al. 2009). Lower biomass production, presumably due to boron deficiency, is observed in mutants (Takano et al. 2006). Aquaporins are proteins with a series of repeated helical domains that form a pore in the cellular membrane and regulate influx and efflux of compounds and regulate a variety of stress responses in plants (Park and Campbell 2015). One role of boron in plant cells is the cross-linking of subunits of rhamnogalacturonan, a cell wall compound (Matoh and Kobayashi 1998, Miwa et al. 2013). This putative gene could contribute to blight resistance by regulating the modification of cell walls during the response to pathogen attack.

A particularly intriguing blight-associated locus, LGA.d, spans a cluster of 6 predicted genes with homology to the wheat leaf rust resistance locus LRK10 (Feuillet et al. 1996). LRK10 is part of a multigene family that is conserved in several grass species (Feuillet and Keller 1999) and *Arabidopsis* (Woo Lim et al. 2015). LRK10-like genes encode membrane-localized receptor kinases. Their extracellular domains are involved in perceiving signals—in many cases, PAMPs or DAMPs—and transmitting that signal to the cytoplasm by phosphorylating another protein. Most of the highly significant SNP associations observed near the LRK10-like gene cluster on Chinese chestnut pseudochromosome LGA are outside of the putative LRK10-like genes; many of them are in AUGUSTUS-predicted transposon genes adjacent to LRK10-like predicted genes. The genes in this cluster show moderate similarity to their closest *Arabidopsis* homologs (30-50% identity) and much stronger similarity to their closest predicted homologs in *Juglans regia* (64-78% identity). Of the three, two (LGA.d.1. and LGA.d.2) are probably expressed in both Chinese and American chestnut – these are more highly conserved, and one is the best predicted protein alignment for two American chestnut cDNA contigs that were expressed more strongly in cankers than in healthy stems. A third AUGUSTUS-predicted gene (LGA.d.3) may actually be an additional exon of the second LRK10-like gene in the cluster (this is what MAKER predicted). Finally, the third or fourth RLK at this locus appears to be fragmentary, but shows evidence of expression in *Cm* and weak evidence of expression in *Cd* (weak because of low identity between transcript and predicted protein). This potentially truncated gene also contains the only non-synonomous associated SNPs within the cluster- the more highly conserved LRK10-like genes in the cluster did not contain any non-synonomous SNPs among the chestnuts we sampled. Some of the SNPs in this predicted gene appear to have severe effects on the polypeptide sequence, including a premature stop codon, major amino acid changes, and a 2 bp insertion. Blight-associated polymorphisms reach their highest frequency in resistant *Cm*, which may indicate that this gene, when functional, heightens blight susceptibility of *Cm*. LGA.d.3 is predicted to contain a single exon that codes a domain similar to the WAK1 (wall-associated kinase 1) c-terminal domain. It could be a cytoplasmic kinase derived from the cytoplasmic portion of an LRK10-like kinase gene or a pseudokinase that retains some role in immune reaction, such as ZED1 in *Arabidopsis,* which is thought to act as a decoy for a fungal receptor (Lewis et al. 2013).

Two predicted genes within the LGA.e locus were the best protein alignments for *Cm* contigs that were expressed at a higher level in cankered versus healthy stem tissue. One had alignments to serine carboxypeptidases in *Arabidopsis* and other plants; the other aligned to a number of predicted genes in plants and animals with a distinctive LisH (lis homology) domain and HEAT repeats, a structure formed from two alpha helices, and an endo/exopolyphosphatase. Since all alignments to this predicted protein aligned to the LisH/HEAT portion or to the polyphosphatase rather than both domains, it seems that the predicted gene may span two actual gene sequences, or incorporate a gene and a pseudogene into one. The LisH/HEAT portion of the predicted gene appears to be highly conserved, with only three non-synonomous SNPs, one of which showed fixation at opposite alleles in American and Chinese chestnut, and which were all confined to a single exon. Most of the highly-significant SNP associations were in the part of the predicted protein with similarity to polyphosphatases. MAKER predicted a slightly longer gene than AUGUSTUS, adding an exon with transcript support from a gene AUGUSTUS predicted as separate. The overall structure of the LisH/HEAT domains consists of two extracellular domains (LisH and HEAT) with two intervening transmembrane domains and a very short (19 amino acids) cytoplasmic domain, as predicted by InterPro. The most similar protein to this locus in Arabidopsis has predicted interactions with an NBS-LRR disease-resistance gene, phospholipid transporters, ARM repeat proteins, and an E3 ubiquitin ligase. This gene may not be directly involved in sensing pathogens, but given its predicted associations with NBS-LRR genes it could form part of a protein complex that acts to effect, modulate, or attenuate a disease response.

The serine carboxypeptidase-like gene (LGA.e.2) did not contain highly associated non-synonomous SNPs, but it did contain strong associations at synonomous SNP loci; there were only five SNPs in AUGUSTUS-predicted exons of this gene. This indicated that any enhanced resistance this gene confers is most likely due to differences in promoters or flanking sites that either enhance or reduce transcription of the gene. cDNA contigs with strong alignments to this predicted protein do not appear in transcriptome datasets for any species other than Chinese chestnut.

The predicted NBS-LRR gene (LGB.a.2) at the LGB.a locus is similar to TAO1, a disease resistance gene in Arabidopsis that is involved in a gene-for-gene resistance relationship with an effector from *Pseudomonas syringae* (Eitas et al. 2008). The predicted gene had no support from available chestnut transcript data, and the very large number of nonsynonymous SNPs in the predicted gene sequence (105; Table 9) indicates that it may in fact be a pseudogene. It is similar (66% protein sequence identity) to a gene from pepper (*Capsicum annuum*). It may be that the predicted gene included part of a true gene and part of a pseudogene, but if that were the case some transcriptome alignments would still be expected. The other potential disease resistance gene at this locus, LGB.a.1, may also be a pseudogene; it had no support from transcriptome data, but the number of non-synonymous SNPs was not as extreme as in LGB.a.2.

The locus LGB.b was identified based on large numbers of highly associated SNPs in a region, not on the number of highly-associated SNPs in a given predicted gene sequence. The predicted genes at this locus show differentiation between *Cm* and *Cd*

(F_{ST} = 0.35-0.65), but none between the most highly resistant *Cm* and less-resistant *Cm*. One, a PIN-LIKES 5-like gene, aligned to a differentially expressed contig from the *Cm* cDNA data. Several non-synonomous SNPs in this predicted gene were fixed for a reference allele in *Cm* and an alternate in *Cd*. PIN-LIKES 5 in Arabidopsis is a gene involved in auxin signalling (Barbez et al. 2012). Auxin is normally associated with the regulation of plant growth, but auxin-signalling pathways have also been implicated in the mediation of disease resistance (Eshraghi et al. 2014), acting in conjunction with jasmonic acid signalling in the resistance reaction to the necrotroph *Alternaria* in *Arabidopsis* (Qi et al. 2012). Its predicted molecular associations are with C2H2-type zinc figner proteins and auxin responsive proteins such as IAA2 and SAUR. It also potentially interacts with a caspase-4 protein that is involved in regulation of programmed cell death. Given its similarity to PIN auxin efflux proteins it could be involved with the trafficking protein particle complex protein LGA.a.3.

The second interesting gene in this region has one highly associated nonsynonymous SNP and several others that are fixed for distinct alleles in American and Chinese chestnuts. This, combined with its homology to a cytochrome P450 oxidase involved in brassinosteroid and jasmonic acid signalling in *Arabidopsis*, make it a reasonable candidate for a role in chestnut blight resistance. Specifically, its predicted role is the synthesis of brassinosteroid hormones, which function in bioitic stress tolerance in several plant/pathogen systems (Nakashita et al. 2003). While it showed no evidence of significant differential expression in response to blight inoculation, it was the best protein alignment for transcripts from both *Cm* and *Cd*.

A cluster of MATE (Multidrug And Toxic compound Extrusion)-like predicted proteins, which are most similar to the DETOXIFICATION-26 and -27 proteins of *Arabidopsis,* were the only blight-associated predicted genes identified in both the cbr scaffold sequences and the whole-genome assembly (LGB.c). It is possible that scaffolds containing crucial genes were not included among the cbr1 scaffolds, and that those genes are represented in some of the other blight-associated regions on LGB. In any case, one of the four predicted (AUGUSTUS) DTX27-like genes in this cluster was also the top candidate gene from the independent analysis of the cbr1 scaffolds. MATE-like genes are involved in the efflux of a wide variety of compounds, including flavonoids and other plant secondary compounds, metals, and organic compounds intended to bind toxic metals, from plant cells. Some MATE genes have been associated with disease-resistance responses, others in auxin signalling, and still others in stress-related ABA signalling (Zhang et al. 2014). If one of these predicted proteins has a role in chestnut blight resistance, it could be in hormone signalling and coordination of the defensive response, or it could be more directly involved as a pump for anti-fungal compounds out of the cell. None of the genes in this cluster had strong alignments to transcripts from any chestnut species, but the high level of seqeunce similarity (64-75% similarity to *Arabidopsis* DTX27, and 85% to predicted proteins from *Malus* and *Vitis*) to sequences from other plant genomes makes it seem unlikely that all the predicted genes in this cluster are pseudogenes An additional predicted gene near the DTX-like cluster, which shows similarity to putative clathrin assembly proteins, is strongly supported by available cDNA data. Clathrin is a protein that forms receptor-associated pits on cell surfaces, and the clathrin-assembly protein in Arabidopsis that aligns to the predicted chestnut protein appears to interact with clathrin-related proteins that have a role in PIN-protein mediated auxin signalling (Kitakura et al. 2011).

LGB.d includes a smaller number of associated SNPs that the others on LGA and LGB, but it includes some potentially biologically relevant genes. There is a predicted 26S proteasome non-ATPase regulatory subunit 4 homolog (PSMD4-like) that aligns to an American chestnut blight canker DEG contig. The *Arabidopsis* protein corresponding to the predicted chestnut protein functions in regulating ABA signalling, senescence, and stress responses (Smalle et al. 2003). This predicted gene appears to be conserved among chestnuts; there were no polymorphisms in any of its predicted exons. Adjacent to the predicted proteasome subunit is a pair of predicted Early Responsive to Dehydration 7 (ERD7)-genes, which are similar to proteins from *Arabidopsis* that are up-regulated during dehydration stress, and in response to ABA (Kyosue et al. 1994). These are small proteins similar to heat-shock proteins. Both had some SNPs in predicted exons, including several non-synonomous SNPs, and strong support from transcripts. Both aligned to transcripts from *Cc* and *Cs,* which indicated that these genes may be involved in general disease resistance responses. The highly associated non-synonomous SNPs in one of the genes segregate between *Cm* and *Cd*, while the second gene had no highly-associated non-synonymous SNPs.

The most likely blight resistance candidate gene in LGB.e is an EDR1-like predicted serine/threonine protein kinase, which aligned to a cDNA contig from American chestnut that showed increased transcription in blight-infected stems. The EDR1 kinase in *Arabidopsis* is a mitogen-activated protein kinase kinase kinase (MAPKKK) that regulates a downstream cascade of mitogen-activated protein kinases (Tang et al. 2002, Tang et al. 2005, Christiansen et al. 2011). It is involved in the regulation of stress response through salycilic acid, ABA, and ethylene signalling pathways. It is also involved in programmed cell death. This predicted gene could play a role in the signalling cascades that initiate resistant or susceptible responses to chestnut blight infection.

This locus also contained a cluster of nine F-box domain-containing predicted genes, similar to FB311 and SKIP-23-like proteins from *Arabidopsis*. *Arabidopsis* F-box proteins are involved in protein ubquination and degradation via the proteosome (Ho et al. 2006), so the function of the predicted F-box genes at this locus in chestnut could be connected to that of the 26s proteosome subunit gene at the LGB.d locus. Two of the F-box proteins have direct support from *Cd* and *Cm* transcriptomes, but most of the associated SNPs in this region are found in other predicted genes without clear transcript support.

The single region on LGC with >100 significantly associated polymorphisms contained several predicted genes with good evidence of transcription in *Cm*, but not in *Cd*. These included two MLP-like proteins, two pentatricopeptide repeat-containing proteins, a KEG E3 ubiquitin-protein ligase, and a LRK10-like leaf rust resistance gene. Of these, all aligned to *Cm* transcripts, but only one to a *Cd* transcript. These genes all appear to be non-polymorphic in their coding sequences, with only a few non-synonymous SNPs. The most likely blight resistance candidate genes were the MLP (Major Latex Protein)-like proteins, which have strong similarity to MLP-like protein 328 in *Arabidopsis*. Neither appeared to be transcribed in American chestnut, and one of the genes contained several nonsynonymous SNPs with one allele in all Chinese chestnuts and another in American chestnuts. MLP-like protein 328 in Arabidopsis is believed to be involved in the plant defensive response, as are other MLP proteins (e.g. Chen and Dai 2010), and is predicted to have molecular interactions with GDSL esterase-lipase, a WD40 domain-containing protein, and TBL38 (Trichome Birefringence-Like 38). Notably, MLP-like proteins are similar to the BetV1 pathogenesis-related / pollen-allergen proteins found on LGJ.

The LGD.a blight-associated locus contains in ELF3 (EARLY-FLOWERING E)-like predicted protein that aligned to a blight DEG contig from Amerrican chestnut, and four other cDNA contigs from Chinese chestnut. It had two non-synonymous SNPs in predicted exons with diffferent alleles fixed in American and Chinese chestnut. Next to this predicted gene was a predicted homolog of HEADING DATE 3B from rice, a predicted cationic peroxidase, a clathrin-interacting protein, and several DETOXIFICATION-like proteins. The cationic peroxidase gene also contained several nonsynonymous SNPs that segregated between *Cm* and *Cd*. In *Arabidopsis*, ELF3 is involved in regulating the initiation of flowering, but is also induced during the plant's response to nematode parasitism. It is up-regulated by auxin and down-regulated by ABA (Liu et al. 2001). The peroxidase is also an attractive candidate for a role in blight resistance, given the documented increase in peroxidase activity during infection, and the role of reactive oxygen species in chestnut blight infection (Hebard et al. 1984). The peroxidase predicted in LGD.a was most similar to PNC1 from *Arachis hypogea*, which is involved in oxidation of toxic reductants, lignin biosynthesis, suberization, and auxin catabolism. All of these processes are likely crucial for a successful defense response against chestnut blight. Finally, the clathrin-interacting protein had a number of highly-associated SNPs within its predicted exons, and several non-synonomous SNPs that segregated between species. It is similar to EPSIN 1 from *Arabidopsis*, which has some role in the transport of clathrin-coated vesicles to the vacuole (Song et al. 2006).

The most interesting predicted genes at locus LGD.b were a probable pectinesterase/pectinesterase inhibitor and a pair of cysteine-rich receptor-like protein kinases (CRKs). The pectinesterase inhibitor only had evidence of expression from *Cm*, but one of the CRKs aligned to transcripts from all four chestnut species. The other predicted CRK-like gene (LGD.b.2) contained a number of non-synonymous SNPs associated with blight resistance (Table 8). Given the importance of dissolving cell walls to the success of a chestnut blight infection, a pectinesterase inhibtor seems like a reasonable candidate for blight resistance, especially since pectin-modifying enzymes have been implicated in responses to other necrotrophic plant diseases and general biotic stress (Atkinson et al. 2013, Nafisi et al. 2015).

LGE.a contained a large number of significantly associated SNPs at a low p-value cutoff, but the genes underlying it remain elusive. The predicted genes (i.e. LGE.a.1 and LGE.a.2) that contain most of the associated SNPs have poor or no homology to known proteins from model plants and are not supported by transcript data from any chestnut species. LGE.a.2 had no alignments to the UniProt or NCBI nr databases (Table 8, Table 9) but a structure prediction on the predicted polypeptide using InterPro revealed a potential transmembrane domain, indicating that this gene could encode a cell-surface protein, but its lack of support from cDNA data makes it unconvincing. Conversely, LGE.a.3, a predicted GDSL esterase-lipase, has strong support from available transcriptome data, but no statistically associated SNPs within the actual gene sequence. It is possible that some of the predicted genes containing associated SNPs at this locus were the result of gene predictions in transposable elements that influence transcription of the esterase-lipase or other genes.

Two genes at LGF.a aligned to periodic tryptophan 2 proteins from yeasts. Similar genes are found in *Arabidopsis*; these are involved in ribosomal assembly and interact with a large number of WD40-domain proteins. One of the pair aligned to transcripts from *Cm*, while the other was only found in *Cd*. Neither contains any associated nonsynonymous SNPs, so whatever nucleotide changes underlie the associations around these genes do not appear to act by causing changes to protein structure. Tryptophan is a precursor for auxin as well as indole glucosinolate compounds that are frequently involved in disease resistance (Denance et al. 2013), so the periodic tryptophan proteins predicted at LGF.a could be involved in furnishing substrates for the generation of auxin or antifungal compounds.

LGF.b contained a predicted serine-threonine protein phosphatase that included a large number of blight resistance-associated polymorphisms (LGF.b.1), but this predicted gene was not supported by available transcriptomic evidence. The nearest predicted gene downstream of the protein phosphatase had a predicted protein sequence similar to the

SKIP23 F-box protein of *Arabidopsis*. This protein is part of a protein ubiquination pathway, so it could be involved in the managing transcription factors and other signalling proteins involved in the blight resistance response.

A conserved ABC (ATP-binding cassette) transporter protein with one statistically associated nonsynonmous SNP in a predicted exon was found within the LGG.a locus. This predicted gene was supported by a transcript from Chinese chestnut, but none from other species. ABC transporters form an enormous, ancient gene family, found in all living organisms, that moves compounds across the cell membrane (Kang et al. 2011). They may transport toxic compounds, surface lipids, and hormones. ABC transporters have been directly implicated in disease resistance, conferring durable disease resistance in wheat (Krattinger et al. 2009), nonhost resistance in *Arabidopsis* (Stein et al. 2006), and regulating hypersensitive cell death in *Arabidopsis* (Kobae et al. 2006). Although the predicted ABC transporter at LGG.a was not particularly homologous to the proteins identified in the latter studies, it could have a similar function in governing the disease resistance response.

A predicted nicotinamidase that appeared to be differentially expressed in Chinese chestnut was found at LGG.b, although the gene itself did not contain any SNPs statistically associated with blight resistance. This gene had evidence of transcription in American chestnut, but no evidence of up-regulation in stem cankers. The most similar nicotinamidase in *Arabidopsis* was involved in ABA signaling; mutants at the NIC1 gene in *Arabidopsis* displayed increased sensitivity to ABA (Wang and Pichersky 2007). NIC1 of *Arabidopsis* is predicted to have interactions with *Arabidopsis* MATE efflux family proteins by the STRING protein interaction database. The chestnut gene LGG.b.1, similar to NIC1, is most likely involved in regulating responses to ABA during reactions to chestnut blight.

LGG.c contained two predicted carboxylesterase genes, one of which was upregulated in canker-infected stems of Chinese chestnut. There were no associated SNPs within the genes, but some were located close to the transcription start site of one CXE gene, indicating that the associated polymorphisms might affect the 5’ regulatory region of the gene rather than the gene itself. The predicted gene (LGG.c.2) that was differentially expressed in Chinese chestnut was similar to CXE5 of Arabidopsis. In

Arabidopsis this protein is predicted to interact with two ribosomal proteins, L21 and L4. Other carboxylesterase genes have been identified as suppressors of disease resistance response to the bacterium *Xanthomonas* (Cunnac et al. 2007). The carboxylesterase-5 gene in *Arabidopsis* is predicted to be a hydrolase of carboxylic esters; it seems likely that this gene's function in chestnut is related to the synthesis or degradation of suberin or other lipid-based compounds that are involved in the defense response. In particular, it could be involved in formation of the phellem layer that contains *Cryphonectria parasitica* in resistant trees.

LGG.d did not contain as many associated SNPs as the other loci on LGG, but it was included here because the predicted genes it contained have transcription profiles and predicted functions that make them convincing blight resistance candidates. In particular, this region contained three ethylene-responsive transcription factors (LGG.d.3-5), one of which is a DEG in American chestnut cankered stems, and also in roots of Japanese chestnut affected by *Phytophthora cinammomi*. *Phytoophthora* is a hemibiotroph, but since it does most damage during its necrotrophic phase the resistance responses to both pathogens most likely involve some common pathways involved in the response to necrotizing pathogens. On the other hand, this gene could be part of a programmed-cell-death response to biotrophs that is exploited by chestnut blight. The fact that one of the ERFs is differentially expressed in American chestnut, but not resistant Chinese chestnut, indicates that elevated expression of this gene might lead to a susceptible response to chestnut blight, but it is also differentially expressed in Japanese chestnut's resistant reaction to *Phytophthora* . Perhaps this ERF is involved in executing a defensive response that is effective against *Phytophthora* but not against *Cryphonectria*. Presumably, other elements are necessary for *Phytophthora* resistance, since American chestnut is susceptible to this pathogen as well as to chestnut blight.

In addition to the ERF genes, LGG.d.1 and LGG.d.2 are predicted serine-threonine protein kinases with strong homology to an *Arabidopsis* gene (PBL27) that is involved in disease responses. Specifically, PBL27 is a cytoplasmic kinase involved in a MAPK (mitogen-activated protein kinase) signalling cascade that responds to the detection of chitin monomers. PBL27 is not a chitin-recognition protein but rather a messenger between the cell surface receptor and other kinases inside the cell that execute the defensive response (Yamada et al. 2016). MAPK cascades are signalling pathways that influence resistance in a variety of plant pathogens, and are part of the early response to the detection of fungal invasion (Meng and Zhang 2013). Neither of the PBL27-like genes predicted at LGG.d showed evidence of differential expression in canker tissue in American chestnut or Chinese chestnut, but it appears that one was supported by transcripts in American, but not Chinese chestnut. Both appear to be conserved, with predicted protein sequences 80% identical to Arabidopsis PBL27 and 90% identical to predicted kinases from *Juglans regia.* All the putative resistance genes in this region are highly conserved, with very few non-synonymous SNPs in any of the ERF or PBL27 predicted genes, so the polymorphisms underlying the statistical associations in this region probably involve regulatory regions rather than the protein sequences of genes.

The most convincing candidate genes at LGJ.a were a pair of predicted genes similar to major pollen allergen proteins, one with strong homology to Pruar 1 of apricot (*Prunus armeniaca*) (LGJ.a.1) and the other to Betv 1 of European white birch (*Betula pendula*) (LGJ.a.2). These genes are homologous, and both cause allergic reactions in humans but serve as pathogenesis-related (PR) proteins in plants, specifically, the PR-10 family (Hoffmann-Sommergruber 2002). PR-10 genes have been implicated in resistant reactions to a wide variety of pathogens, including witch's broom of cacao, caused by a hemibiotroph (Menezes et al. 2012), *Cochliobolus* and *Colletotrichum* in sorghum (Lo et al. 1999), *Magnaporthe grisea* in rice (McGee et al. 2001) and even the parasitic plant dodder in alfalfa (Borsics and Lados 2002). They tend to be expressed most strongly in the immediate area of pathogen attack (Lo et al. 1999, McGee et al. 2001). Some are expressed constitutively, however, and their precise molecular function remains unknown (Fernandes et al. 2013). One of the two predicted chestnut genes on LGJ (LGJ.a.1) showed evidence of expression in all four chestnut species, but no significant differential expression in diseased tissue. The protein sequences themselves are highly conserved with no nonsynonymous SNPs. Both genes had highly associated SNPs directly upstream. Tajima's D values indicate that LGJ.a.1 has been subject to purifying selection in American chestnut (D = -1.22) but not in resistant Chinese chestnuts (D = 1.79), and F_{ST} among species indicated a high level of interspecific divergence for both genes. These proteins are most likely involved in a conserved, general disease response.

LGK.a contained a predicted gene that was similar to MIEL1 of *Arabidopsis,* an E3 ubiquitin-protein ligase that serves as a regulator of cell death and defense against pathogens. Specifically, it regulates cell death by marking for destruction (ubquinating) a transcription factor, MYB30, that positively regulates the hypersensitive disease response, sparing plants the expense of carrying out a disease response when it is not necessary (Marino et al. 2013). This predicted gene aligned to a transcript that is expressed more in cankered stems than healthy stems of Chinese chestnut and a transcript from Japanese chestnut, but none from American or European chestnut. Since this indicates that the resistant tree is expressing a gene that attenuates the defensive reaction, it seems likely that down-regulation of certain defense responses, in particular the HR, could be important to resisting *Cryphonectria parasitica* infection. The runaway cell death seen in American chestnut could be the result of a failure to down-regulate HR responses in the presence of chestnut blight.

More statistically-associated SNPs, and more associated SNPs within genes, were found on the LGL.a locus than on any other linkage group. The vast majority of these SNPs were in predicted retroelement-associated genes and genes with no homology to known proteins and no support from the available chestnut transcriptomes. The most convincing candidate genes in this region were two predicted MAIL3-like serine/threonine protein phosphatase homologs, which both contained some associated non-synonymous polymorphisms, but were not supported by any available transcript data. The most similar protein phosphatase to these predicted genes in *Arabidopsis* may be involved in managing cell division, organization and growth in meristem tissues (Wenig et al. 2013). Other serine-threonine protein phosphatases are directly involved in disease resistance, as they interact with NBS-LRR disease resistance proteins (van Bentem et al. 2005).

LGL.b was most notable for a predicted PBL4-like serine-threonine protein kinase (LGL.b.2) that appears to be expressed in Chinese chestnut, but not in other species. Similar to the PBL27-like genes of the LGG.d locus, this kinase is most likely involved in signalling a MAPK cascade in the disease resistance reaction. It could be involved in fine-tuning the Chinese chestnut defensive response, and if it is truly not expressed in

American chestnut, susceptibility in that species could be partly due to lacking the regulation of disease resistance pathways that this predicted gene may provide.

We identified a large cluster (10 genes or more) of predicted disease resistance genes with NBS-LRR and RPW8 domains at LGL.c. In addition to their homology to other R genes, genes in this cluster resembled classic R genes in several respects: their occurrence in a cluster, relatively high nucleotide diversity, and generally high Tajima's D values (up to 2.5 in Chinese chestnuts) indicating a history of diversifying selection. A number of these genes showed differential expression in response to pathogens. LGL.c.2 was up-regulated in American chestnut inoculated with chestnut blight as well as Japanese and European chestnut roots inoculated with *Phytophthora cinammomi*. Two additional predicted R genes in the cluster, LGL.c.5 and LGL.c.8, were also differentially expressed in European chestnut roots affected by *Phytophthora*. LGL.c.8 contains more associated nonsynonymous SNPs than the other genes in the cluster, but did not show evidence of expression in Chinese chestnut. NBS-LRR genes are often involved in signalling the HR and programmed cell death, so the increased expression of genes at this locus in susceptible chestnut species could indicate a cell-death response that is exploited by the necrotroph *Cryphonectria parasitica* and perhaps by *Phytophthora* in its necrotrophic phase. The NBS-LRR genes in this cluster contain an RPW8-like domain. RPW8 is a small protein that is involved in resistance to biotrophs, specifically powdery mildews, in Arabidopsis. It initiates HR lesions to contain infections and leads to the induction of PR genes (Wang et al. 2007). Notably, overexpression of RPW8 in Arabidopsis increases susceptibility to necrotrophic pathogens (Wang et al. 2007).

A DEG from American chestnut cankers aligns to a predicted retroelement-related protein at LGL.d, but sequence identity is fairly low (50%) so this alignment may be spurious. The transposable element-like gene appears to be highly conserved at the DNA level, however, with a large number of SNPs in predicted introns and none in predicted exons. Several predicted eukaryotic peptide chain release factor subunit genes (ERF1Z; LGL.d.2-4) with resistance-associated nonsynonymous SNPs (LGL.d.2,3) were near the predicted transposon-like gene. Transgenic *Arabidopsis* plants with suppressed expression of ERF1 showed increased lignification of phloem tissue and some unusual growth phenotypes (Petsch et al. 2005). ERF proteins interact with ribosomal proteins to terminate the translation of polypeptides. There are three well-documented ERF genes in Arabidopsis (Chapman and Brown 2004); this cluster may represent an additional expansion of the gene family in chestnut. One of the three copies predicted here was highly conserved (LGL.d.4) with no nonsynonymous SNPs. All three showed a high level of conservation (>90% sequence identity) with predicted proteins from other plant genomes, and all three had some transcriptome support from chestnut transcriptomes. Given their inferred role in cell growth and division, it seems likely that these genes are involved in lignification and callus formation rather than innate immune responses.

3. 5 Conclusions

The molecular basis of resistance to necrotrophic pathogens in plants is generally complex, and it appears that chestnut blight is no exception. Our association analysis, while it included a relatively small sample of individual phenotypes, revealed patterns of association that correspond with previous observations of blight resistance QTL on linkage groups B, F, and G while also identifying loci on other linkage groups that appear to be associated with resistance. The genes in these regions included some that resemble classical R genes (LRK10-like receptor kinases and NBS-LRR genes) while others are involved in hormone signaling, kinase cascades, and cell-wall modification. Several of these genes correspond to differentially-expressed cDNA contigs from a previous transcriptome experiment. *Cryphonectria parasitica* may be so virulent on American chestnut because it exploits pattern-recognition receptors and the hypersensitive programmed-cell-death response they initiate to cause rampant, uncontained cankers. Chinese chestnut may be more resistant because it successfully manipulates these disease-resistance reactions to be less severe, so that the pathogen can be contained. While the receptor-like loci seem to have unique genotypes and high polymorphism in the most resistant Chinese chestnuts, they do not seem to be the loci underlying species differences in resistance, which implies that genes downstream in the signalling pathway are sufficient to confer some level of resistance in American chestnut. High nucleotide diversity at some resistance genes seems to confer an advantage in the most resistant Chinese chestnut, but some of the candidate genes we identified essentially had a fixed genotype with low heterozygosity in all Chinese chestnuts. Therefore, it may be that a limited number of Chinese chestnut resistance donors would not be a major impediment to introgressing durable resistance into American chestnut. However, including diverse alleles from the “best” resistant Chinese chestnuts at those genes where diversity confers an advantage might produce the most resistant offspring. These candidate genes require a great deal of additional work to validate their role in blight resistance, but the work described here should provide a strong starting point for those research efforts.

3.6 Literature cited

Agrios GN (2005) Plant Pathology: 5th Edition. Elsevier Academic Press, Burlington, MA, USA ISBN-13 978-0-12-044565-3.

Anagnostakis SL (1987) Chestnut blight: the classical problem of an introduced pathogen. Mycologia 79(1): 23-27.

Anagnostakis SL (1992) Chestnut bark tannin assays and growth of chestnut blight fungus on extracted tannin. Journal of Chemical Ecology 18(8):1365-1373.

Anagnostakis SL (2001). The effect of multiple importations of pests and pathogens on a native tree. Biological Invasions 3: 245-254.

Atkinson NG, Lilley CJ, Urwin PE (2013) Identification of genes involved in the response of *Arabidopsis* to simultaneous biotic and abiotic stresses. Plant Phys 162(4):2028-2041.

Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A Genome-Wide Survey of R Gene Polymorphisms in *Arabidopsis*. Plant Cell 18(8): 1803-1818.

Bakker EG, Traw MB, Toomajian C, Kreitman M, Bergelson J (2008) Low levels of polymorphism in genes that control the activation of defense response in *Arabidopsis thaliana* Genetics 178(4):2031-2043.

Barakat, A., DiLoreto, D.S., Zhang, Yi, Smith, C., Baier, K., Powell, W.A., Wheeler, N., Sederoff, R., Carlson, J.E. 2009. Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (C. mollissima) in response to the chestnut blight infection. BMC Plant Biology 9:51.

Barakat A, Staton M, Cheng C-H et al. (2012) Chestnut resistance to the blight disease: insights from transcriptome analaysis. BMC Plant Biology 12:38.

Barbez E, Kubes M, Rolcik J et al. (2012) A novel putative auxin carrier family regulates intracellular auxin homeostasis in plants. Nature 485:119-122.

Beckers GJM, Jaskiewicz M, Liu Y, Underwood WR, He SY, Zhang S, Conrath U (2009) Mitogen-activated protein kinases 3 and 6 are required for full priming of stress responses in *Arabidopsis thaliana*. Plant Cell 21:944-953.

Bent AF, Mackey D (2007) Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions. Annual Review of Phytopathology 45:399-436.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, btu170.

Boller T, Yang He S (2009) Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. Science 324(5928): 742-744.

Borsics T, Lados M (2002) Dodder infection induces the expression of a pathogenesis-related gene of the family PR-10 in alfalfa. J Exp Bot 53(375):1831-1835.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. Methods Mol Biol 1374:23-54.

Buchfink B, Xie C, Huson D (2015) Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59-60.

Burnham, C.R., Rutter, P.A., French, D.W. 1986. Breeding Blight-Resistant Chestnuts. Plant Breeding Reviews 4: 347-397.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18(1):188-196.

Cessna SG, Sears VE, Dickman MB, Low PS (2000) Oxalic acid, a pathogenicity factor for *Sclerotinia sclerotiorum,* suppresses the oxidative burst of the host plant. Plant Cell 12(11):2191-2199.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

Chapman B, Brown C (2004) Translation termination in *Arabidopsis thaliana*: characterization of three versions of release factor 1. Gene 341:219-225.

Chassot C, Nawrath C, Metraux J-P (2007) Cuticular defects lead to full immunity to a major plant pathogen. Plant Journal 49:972-980.

Chen J-Y, Dai X-F (2010) Cloning and characterization of the *Gossypium hirsutum* major latex protein and functional analysis in *Arabidopsis thaliana.* Planta 231(4):861-873.

Chow BY, Helfer A, Nusinow DA, Kay SA (2012) ELF3 recruitment to the PRR9 promoter requires other Evening Complex members in the *Arabidopsis* circadian clock. Plant Signaling and Behavior 7:170-173.

Christiansen KM, Gu Y, Rodibaugh N, Innes RW (2011) Negative regulation of defence signaling pathways by the EDR1 protein kinase. Molecular Plant Pathology 12:746-758.

Clarke JD, Volko SM, Ledford H, Ausubel FM, Dong X (2000) Roles of salicylic acid, jasmonic acid, and ethylene in *cpr*-induced resistance in *Arabidopsis.* The Plant Cell 12(11): 2175-2190.

Cui X, Cao X (2014) Epigenetic regulation and functional exaptation of transposable elements in higher plants. Current Opinion in Plant Biology 21:83-88.

Cunnac S, Wilson A, Nuwer J, Kirik A, Baranage G, Mudgett MB (2007) A conserved carboxylesterase is a SUPRESSOR OF AVBRST-ELICITED RESISTANCE in *Arabidopsis*. The Plant Cell 19:688-705.

Danecek P, Auton A, Abecasis G et al. (2011) The Variant Call Format and VCFtools. Bioinformatics 27(15):2156-2158.

Danquah A, de Zelicourt A, Colcombet J, Hirt H (2014) The role of ABA and MAPK signaling pathways in plant abiotic stress responses. Biotechnology Advances 32:40-52.

Denance N, Sanchez-Vallet A, Goffner D, Molina A (2013) Disease resistance or growth: the role of plant hormones in balancing immune responses and fitness costs. Frontiers in Plant Science 4:155.

DePristo M, Banks E, Garimella K et al. (2011) A framework for variation discovery and genotyi,./ping using next-generation DNA sequencing data. Nature Genetics 43:491-498.

DiDonato RJ, Arbuckle E, Buker S, Sheets J, Tobar J, Totong R, Grisafi P, Fink GR, Celenza JL (2003) *Arabidopsis* ALF4 encodes a nuclear-localized protein required for lateral root formation. The Plant Journal 37(3):340-353.

Diener AC, Gaxiola RA, Fink GR (2001) *Arabidopsis* ALF5, a multidrug efflux transporter gene member, confers resistance to toxins. Plant Cell 13:1625-1637.

Dobon A, Canet JV, Garcia-Andrade J, Angulo C, Neumetzler L, Persson S, Vera P (2015) Novel disease susceptibility factors for fungal necrotrophic pathogens in Arabidopsis. PLOS Pathogens https://doi.org/10.1371/journal.ppat.1004800

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemistry Bulletin 19:11-15.

Eitas TK, Nimchuk ZL, Dangl JL (2008) Arabidopsis TAO1 is a TIR-NB-LRR protein that contributes to disease resistance induced by the *Pseudomonas syringae* effector AvrB. Proceedings of the National Academy of Sciences USA 205:6475-6480.

Eshraghi L, Anderson JP, Aryamanesh N, McComb JA, Shearer B, St J E Hardy G (2014) Suppression of the auxin response pathway enhances susceptibility to *Phytophthora cinammomi* while phosphate-mesiated resistance stimulates the auxin signaling pathway. BMC Plant Biology 14:68 doi:10.1186/1471-2229-14-68.

Fang, G.-C., Blackmon, B.P., Staton, M.E., Nelson, C.D., Kubisiak, T.L. et al. 2013. A physical map of the Chinese chestnut genome and its integration with the genetic map. Tree Genetics and Genomes 9:535-537.

Faris JD, Zhang Z, Lu H et al. (2010) A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. Proceedings of the National Academy of Sciences USA 107(30):13544-13549.

Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. Proceedings of the National Academy of Sciences USA 96:8265-8270.

Feuillet C, Schlachermayr G, Keller B (1997) Molecular cloning of a new receptor-like kinase gene encoded at the Lr10 disease resistance locus of wheat. The Plant Journal 11(1):45-52.

Fernandes H, Michalska K, Sikorski M, Jaskolski M (2013) Structural and functional aspects of PR-10 proteins. FEBS Journal 280:1169-1199.

Friesen TL, Faris JD, Solomon PS, Oliver RP (2008) Host-specific toxins: effectors of necrotrophic pathogenicity. Cell Microbiology 10(7): 1421-8.

Fry W (2008) *Phytophthora infestans*: the plant (and R gene) destroyer. Molecular Plant Pathology 9(3):385-402.

Frye CA, Innes RW (1998) An *Arabidopsis* mutant with enhanced resistance to powdery mildew. Plant Cell 110:947-956.

Frye CA, Tang D, Innes RW (2001) Negative regulation of defense responses in plants by a conserved MAPKK kinase. Proceedings of the National Academy of Sciences USA 98:373-378.

Gu Y, Innes RW (2011) The KEEP ON GOING protein of *Arabidopsis* recruits the ENHANCED DISEASE RESISTANCE 1 protein to trans-Golgi network/ early endosome vesicles. Plant Physiology 155:1827-1838.

Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W.J. et al. 2013. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nature Genetics 45(1):51.

Hayashi K, Yoshida H (2008) Refunctionalization of the ancient rice blast disease resistance gene Pit by the recruitment of a retrotransposon as a promoter. The Plant Journal 57:413-425.

Hayden KJ, Garbeletto M, Knaus BJ, Cronn RC, Rai H, Wright JW (2014) Dual RNA-seq of the plant pathogen *Phytophthora ramorum* and its tanoak host. Tree Genetics and Genomes 10(3):489-502.

Hebard, FV (2006) The backcross breeding program of the American Chestnut Foundation. pp. 61-77. In: K.C. Steiner and J.E. Carlson (eds.), Restoration of the American chestnut tree to forest lands- proceedings of a conference and workshop. May 4-6 2004, North Carolina Arboretum. Natural Resources Rep. NPS/NCR/CUE/NRR- 2006/001, National Park Service, Washington, D.C.

Hebard FV, Griffin GJ, Elkins JR (1984) Developmental histopathology of canker incited by hypovirulent and virulent isolates of *Endothia parasitica* on susceptible and resistance chestnut trees. Phytopathology 74: 140-149.

Herrero E, Kolmos E, Bujdoso N (2012) EARLY FLOWERING 4 recruitment of EARLY FLOWERING 4 in the nucleus sustains the *Arabidopsis* circadian clock. Plant Cell 24:428-443.

Hiruma K, Nishiuchi T, Kato T, Bednarek P, Okuno T, Schulze-Lefert P, Takano Y (2011) *Arabidopsis* ENHANCED DISEASE RESISTANCE 1 is required for pathogen-induced expression of plant defensins in nonhost resistance, and acts through interference of MYC2-mediated repressor function. The Plant Journal 67:980-992.

Ho MS, Tsai PI, Chien CT (2006) F-box proteins: the key to protein degradation. Journal of Biomedical Sciences 13(2):181-91.

Jakobson L, Lindgren LO, Verdier G, Laanemets K, Brosche M, Beisson F, Kollist H (2016) BODYGUARD is required for the biosynthesis of cutin in *Arabidopsis*. New Phytologist 211:614-626.

Jansen RK, Saski C, Lee S-B, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three rosids (*Castanea, Prunus, Theobroma*): Evidence for at least two independent transfers of rpl22 to the nucleus. Molecular Biology and Evolution 28(1):835-847.

Jaynes RA (1974) Genetics of chestnut. USDA Forest Service Research Paper WO-17, 24 p.

Jones JDG, Dangl JL (2006) The plant immune system. Nature 444:323-329.

Laluk K, Mengiste T (2010) Necrotroph attacks on plants: wanton destruction or covert extortion? Arabidopsis Book 8: e0136. doi: 10.1199/tab.0136

Lam, H-M, X, X., Liu, X., Chen, W., Yang, G., et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature Genetics 42: 1053-1059.

Lee TH, Guo H, Wang X, Kim C, Paterson AH (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162.

Lewis JD, Lee AH-Y, Hassan JA et al. (2013) The *Arabidopsis* ZED1 pseudokinase is required for ZAR1-mediated immunity induced by the *Pseudomonas syringae* type III effector HopZ1a. Proceedings of the National Academy of Sciences USA 110(46):18722-18727.

Kamiya T, Tanaka M, Mitani N, Ma JF, Maeshima M, Fujiwara T (2009) NIP1;1, an aquaporin homolog, determines the arsenite sensitivity of *Arabidopsis thaliana*. Journal of Biological Chemistry 284:2113-2120.

Kang J, Park J, Choi H, Burla B, Kretzschmar T, Lee Y, Martinoia E (2011) Plant ABC Transporters. Arabidopsis Book 9:e0153.

Kitakura S, Vanneste S, Robert S, Löfke C, Teichmann T, Tanaka H, Friml J (2011) Clathrin mediates endocytosis and polar distribution of PIN auxin transporters in *Arabidopsis.* Plant Cell 23(5):1920-1931.

Kiyosue T, Yamaguchi-shinozaki K, Shinozaki K (1994) Cloning of cDNAs for genes that are early-responsive to dehydration stress (ERDs) in *Arabidopsis thaliana.* Plant Molecular Biology 25:791-798.

Kobae Y, Sekino T, Yoshioka H, Nakagawa T, Martionoia E, Maeshima M (2006) Loss of AtPDR8, a plasma membrane ABC transporter of *Arabidopsis thaliana*, causes hypersensitive cell death upon pathogen infection. Plant Cell Physiology 47(3):309-318.

Krattinger SG, Lagudah ES, Spielmeyer W et al. (2009) A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. Science 323(5919):1360-1363.

Kubisiak TL, Hebard FV, Nelson CD, Zhang J, Bernatzky R, Huang J, Anagnostakis SL, Doudrick RL (1997) Molecular mapping of resistance to blight in an interspecific cross in the genus *Castanea*. Phytopathology 87:751-759.

Kubisiak TL, Nelson CD, Staton ME, Zhebentyayeva T, Smith C, Olukolu BA, Fang G-C, Hebard FV, Anagnostakis S, Wheeler N, Sisco PH, Abbott AG, Sederoff RR (2013) A transcriptome-based genetic map of Chinese chestnut (*Castanea mollissima*) and identification of regions of segmental homology with peach (*Prunus persica*). Tree Genetics and Genomes 9:557-571.

Kurdyukov S, Faust A, Nawrath C et al. (2006) The epidermis-specific extracellular BODYGUARD controls cuticle development and morphogenesis in *Arabidopsis.* Plant Cell 18:321-339.

Kurepa J, Toh E-A, Smalle JA (2008) 26S proteasome regulatory particle mutants have increased oxidative stress tolerance. The Plant Journal 53:102-114.

Lee S-J, Rose JKC (2010) Mediation of the transition from biotrophy to necrotrophy in hemibiotrophic plant pathogens by secreted effector proteins. Plant Signaling and Behavior 5(6): 769-772.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 35:1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., et al. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Liu XL, Covington MF, Fankhauser C, Chory J, Wagner DR (2001) ELF3 encodes a circadian clock-regulated nuclear protein that functions in an *Arabidopsis* PHYB signal transduction pathway. Plant Cell 13:1293-1304.

Lo S-C, Hipskind JD, Nicholson RL (1999) cDNA cloning of a sorghum pathogenesis-related protein (PR-10) and differential expression of defense-related genes following inoculation with *Coclhiobolus heterostrophus* or *Colletotrichum sublineolum*. Molecular Plant-Microbe Interactions 12(6): 479-489.

Lorang J, Kidarsa T, Bradford CS, Gilbert B, Curtis M, Tzeng S-C, Maier CS, Wolpert TJ (2012) Tricking the guard: exploiting plant defense for disease susceptibility. Science 338(6107):659-662.

Marino D, Froidure S, Canonne J, Ben Khaled S, Khafif M, Pouzet C, Jauneau A, Roby D, Rivas S (2013) Arabidopsis ubiquitin ligase MIEL1 mediates degradation of the transcription factor MYB30 weakening plant defence. Nature Communications 4:1476-1476.

Matoh T, Kobayashi M (1998) Boron and calcium, essential inorganic constituents of pectic polysaccharides in higher plant cell walls. Journal of Plant Research 111(1):179-190.

Mauch-Mani B, Mauch F (2005) The role of abscisic acid in plant-pathogen interactions. Current Opinion in Plant Biology 8:409-414.

Mayer AM, Staples RC, Gil-ad NL (2001) Mechanisms of survival of necrotrophic fungal plant pathogens in hosts expressing the hypersensitive response. Phytochemisty 58(1):33-41.

McCarroll DR, Thor E (1978) Death of a chestnut: the host pathogen interaction. In: Proceedings of the American Chestnut Symposium, Morgantown, West Virginia, USA January 4-5 1978.

McCarroll DR, Thor E (1985) Do “toxins” affect pathogenesis by *Endothia parasitica*? Physiological Plant Pathology 26(3):357-366.

McCarroll DR, Thor E (1985) Pectolytic, cellulytic and proteolytic activities expressed by cultures of *Endothia parasitica*, and inhibition of these activities by components extracted from Chinese and American chestnut inner bark. Physiological Plant Pathology 26(3):367-378.

McGee JD, Hamer JE, Hodges TK (2001) Characterization of a *PR-10* pathogenesis-related gene family induced in rice during infection with *Magnaporthe grisea*. Molecular Plant-Microbe Interactions 14(7):877-886.

McKenna A, Hanna M, Banks E et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation sequencing data. Genome Research 20:1297-1303.

Melnyk CW, Schuster C, Leyser O, Meyerowitx EM (2015) A developmental framework for graft formation and vascular reconnection in *Arabidopsis thaliana.* Current Biology 25(10):1306-1318.

Metcalf H (1912) The chestnut bark disease. p. 363-372 in: "Yearbook of the Department of Agriculture for 1912," Washington, D.C.

Menezes SP, dos Santos JL, Cardoso THS, Pirovani CP, Micehli F, Noronha FSM, Alves AC, Faria AMC, da Silva Gesteira A (2012) Evaluation of the allergenicity potential of TcPR-10 protein from *Theobroma cacao.* https://doi.org/10.1371/journal.pone.0037969

Meng X, Zhang S (2013) MAPK cascades in plant disease resistance signaling. Annu Rev Phytopathology 51:245-66.

Mengiste T (2012) Plant immunity to necrotrophs. Annual Review of Phytopathology 50:267-294.

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Research 8:1113-1130.

Morel J-B, Dangl JL (1997) The hypersensitive response and the induction of cell death in plants. Cell Death & Differentiation 4(8):671-683.

Moricca S, Linaldeddu BT, Ginettri B, Scanu B, Franceschini A, Ragazzi A (2016) Endemic and emerging pathogens threatening cork oak trees: management options for conserving a unique forest ecosystem. Plant Disease 100(11):2184-2193.

Nafisi M, Fimognari L, Sakuragi Y (2015) Interplays between the cell wall and phytohormones in interaction between plants and necrotrophic pathogens. Phytochemistry 112:63-71.

Nakashita H, Yasuda M, Nitta T, Asami T, Fujioka S, Arai Y, Sekimata K, Takatsuto S, Yamaguchi I, Yoshida S (2003) Brassinosteroid functions in a broad range of disease resistance in tobacco and rice. The Plant Journal 33(5):887-98.

Naramoto S, Nodzylski T, Dainobu T, Takatsuka H, Okada T, Friml J, Fukuda H (2014) VAN4 encodes a putative TRS120 that is required for normal cell growth and vein development in *Arabidopsis*. Plant Cell Physiology 55:750-763.

Negi P, Rai AN, Suprasanna P (2016) Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. Frontiers in Plant Science 7:1-20.

Nelson RR (1978) Genetics of horizontal resistance to plant diseases. Annual Review of Phytopathology 16:359-378.

Park WJ, Campbell BT (2015) Aquaporins as targets for stress tolerance in plants: genomic complexity and perspectives. Turkish Journal of Botany 39:879-886.

Parlevliet JE, Zadoks JC (1977) The integrated concept of disease resistance: a new view including horizontal and vertical resistance in plants. Euphytica 26:5.

Petsch KA, Mylne J, Botella JR (2005) Cosuppression of Eukaryotic Release Factor 1-1 in Arabidopsis affects cell elongation and radial cell division. Plant Physiology 139(1):115-126.

Poland JA, Balint-Kurti PJ, Wisser RJ, Pratt RC, Nelson RJ (2009) Shades of gray: the world of quantitative disease resistance. Trends in Plant Science 14(1):21-29.

Puhalla JE, Anagnostakis SL (1970) Genetics and nutritional requirements of *Endothia parasitica*. Phytopathology 61:169-173.

Protein nr database [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] - [cited 2017 May 18]. Available from: https://www.ncbi.nlm.nih.gov/protein/nr

Qi L, Yan J, Li Y et al. (2012) *Arabidopsis thaliana* plants differentially modulate auxin biosynthesis and transport during defense responses to the necrotrophic pathogen *Alternaria brassicicola.* New Phytologist 195(4):872-882.

Qi X, Kaneda M, Chen J, Geitmann A, Sheng H (2011) A specific role for *Arabidopsis* TRAPPII in post-Golgi trafficking that is crucial for cytokinesis and cell polarity. The Plant Journal 68:234-248.

Qin L, Gao X, Cheng J, Liu S (1999) Evaluation of Chinese chestnut cultivars for resistance to Cryphonectria parasitica. Proc. 2nd Int. Symp. on Chestnut, Ed. G. Salesses. Acta Horticulturae 494, ISHS 1999.

Quigley F, Rosenberg JM, Shachar-Hill Y, Bohnert HJ (2002) From genome to function: the *Arabidopsis* aquaporins. Genome Biology 3:1-17.

Rahman TA, Oirdi ME, Gonzalez-Lamothe R, Bouarab K (2012) Necrotrophic pathogens use the salicylic acid signaling pathway to promote disease development in tomato. Molecular Plant Microbe Interactions 25(12):1584-1593.

Remy E, Duque P (2014) Beyond cellular detoxification: a plethora of physiological roles for MDR transporter homomlogs in plants. Frontiers in Physiology 5:201. doi : 10.3389/fphys.2014.00201

Ribas AF, Cenci A, Combes M-C, Etienne J, Lashermes P (2011) Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. BMC Genomics 12:240.

Rose LE, Bittner-Eddy PD, Langley CH, Holub EB, Michelmore RW, Beynon JL (2004) The maintenance of extreme amino acid diversity at the disease resistance gene, RPP13, in *Arabidopsis thaliana.* Genetics 166(3):1517-1527.

Serrazina SMT, Santos C, Machado H, Pesquita C, Vicentini R, Pais M, Sebastiana M, Costa RL (2015) *Castanea* root transcriptome in response to *Phytophthora cinnamomi* challenge. Tree Genetics and Genomes 11(1) DOI: 10.1007/s11295-014-0829-7

Shain L, Gao S (1995) Activity of polygalacturonase produced by Cryphonectria parasitica in chestnut bark and its inhibition by extracts from American and Chinese chestnut. Physiological and Molecular Plant Pathology 46(3): 199-213.

Shi G, Zhang Z, Friesen TL et al. (2016) The hijacking of a receptor kinase-driven pathway by a wheat fungal pathogen leads to disease. Scientific Advances 2:e1600822.

Slavov, G.T., DiFazio, S.P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E. et al. 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. New Phytologist 2012 196:713-725.

Smalle J, Kurepa J, Yang P, Emborg TJ, Babiychuk E, Kushnir S, Vierstra RD (2003) The pleiotropic role of the 26S proteasome subunit RPN10 in *Arabidopsis* growth and development supports a substrate-specific function in abscisic acid signalling. Plant Cell 15:965-980.

Sniezko RA, Smith J, Liu J-J, Hamelin RC (2014) Genetic resistance to fusiform rust in southern pines and white pine blister rust in white pines—a contrasting tale of two rust pathosystems—current status and future prospects. Forests 5:2050-2083.

Song J, Lee MH, Lee G-J, Yoo CM, Hwang I (2006) *Arabidopsis* EPSIN1 plays an important role in vacuolar trafficking of soluble cargo proteins in plant cells via interactions with clathrin, Ap-1, VTI11, and VSR1. Plant Cell 18:2258-2274.

Soriano JM, Joshi SG, van Kaauwen M, Noordijk Y, Growenwold R, Henken B, van de Weg WE, Schouten HJ (2009) Identification and mapping of the novel apple scab resistance gene Vd3. Tree Genetics and Genomes 5(3):475-482.

Sugimoto K, Jiao Y, Meyerowitz EM (2010) *Arabidopsis* regeneration from multiple tissues occurs via a root development pathway. Developmental Cell 18:463-471.

Stanke M, Schoeffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62.

Staton ME, Addo-Quaye C, Cannon N, Tomsho LP, Drautz D, Wagner TK, Zembower N, Ficklin S, Saski C, Burhans R, Schuster SC, Abbott AG, Nelson CD, Hebard FV, Carlson JE (2014) *The Chinese chestnut (Castanea mollissima) genome version 1.1*, http://www.hardwoodgenomics.org/chinese-chestnut-genome, Access date August 2, 2016.

Stein M, Dittgen J, Sanchez-Rodriguez C, Hou B-H, Molina A, Schulze-Lefert P, Lipka V, Somerville S (2006) *Arabidopsis* PEN3/PDR8, an ATP binding cassette transporter, contributes to nonhost resistance to inappropriate pathogens that enter by direct penetration. The Plant Cell 18:731-746.

Swett CL, Kirkpatrick SC, Gordon TR (2016) Evidence for a hemibiotrophic association of the pitch canker pathogen *Fusarium circinatum* with *Pinus radiata.* Plant Disease 100(1): 79-84.

Takano J, Wada M, Ludewig U, Schaaf G, von Wiren N, Fujiwara T (2009) The *Arabidopsis* major intrinsic protein NIP5;1 is essential for efficient boron uptake and plant development under boron limitation. Plant Cell 18:1498-1509.

Tang D, Christiansen KM, Innes RW (2005) Regulation of plant disease resistance, stress response, cell death, and ethylene signaling in *Arabidopsis* by the EDR1 protein kinase. Plant Physiology 138:1018-1026.

Tang D, Innes RW (2002) Overexpression of a kinase-deficient from of the EDR1 gene enhances powdery mildew resistance and ethylene-induced senescence in *Arabidopsis.* The Plant Journal 32:975-983.

Thellmann M, Rybak K, Thiele K, Wanner G, Assaad FF (2010) Tethering factors required for cytokinesis in *Arabidopsis.* Plant Physiology 154:720-723.

Thakur, S., Gupta, Y.K., Singh, P.K., Rathour, R., Variar, M. et al. 2013. Molecular diversity in rice blast resistance gene Pi-ta makes it highly effective against dynamic population of Magnoporthe oryzae. Functional and integrated genomics 13:309-322.

The UniProt consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45: D158-D169.

Van Bentem S, Vossen JH, de Vries KJ, van Wees S, Tameling WIL, Dekker JL, de Koster CG, Haring MA, Takken FLW, Cornelissen BJC (2005) Heat shock protein and its co-chaperone protein phosphatase 5 interact with distinct regions of the tomato I-2 disease resistance protein. The Plant Journal 32(2):284-298.

Van der Auwera GA, Carneiro M, Hartl C et al. (2013) From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43:11.10.1-11.10.33.

Van Der Biezen EA, Jones JDG (1998) Plant disease resistance proteins and the gene-for-gene concept. Trends in Biochemical Sciences 23(12): 454-456

van Hulten M, Pelser M, van Loon LC, Pieterse CM, Ton J (2006) Costs and benefits of priming for defense in *Arabidopsis*. Proceedings of the National Academy of Sciences USA 103:5602-5607.

Van Der Plank JE (1966) Horizontal (polygenic) and vertical (oligogenic) resistance against blight. American Potato Journal 43(2):43-52.

Wang G, Pichersky E (2007) Nicotinamidase participates in the salvage pathway of NAD biosynthesis in Arabidopsis. The Plant Journal 49:1020-1029.

Wang X, Jiang N, Liu J, Liu W, Wang G-L (2014) The role of effectors and host immunity in plant-necrotrophic fungal interactions. Virulence 5(7): 722-732.

Wang W, Devoto A, Turner JG, Xiao S (2007) Expression of the membrane-associated resistance protein RPW8 enahnces basal defense against biotrophic pathogens. Molecular Plant-Microbe Interactions 20(8): 966-976.

Wang Z-Y, Xiong L, Li W, Zhu J-K,Zhu J (2011) The plant cuticle is required for osmotic stress regulation of abscisic acid biosynthesis and osmotic stress tolerance in *Arabidopsis*. Plant Cell 23:1971-1984.

Wawrzynska A, Christiansen KM, Lan Y, Rodibaugh NL, Innes RW (2008) Powdery mildew resistance conferred by loss of the ENHANCED DISEASE RESISTANCE 1 protein kinase is suppressed by a missense mutation in KEEP ON GOING, a regulator of abscisic acid signalling. Plant Physiology 148:1510-1522.

Wei C, Chen J, Kuang H (2016) Dramatic number variation of R genes in Solanaceae species accounted for by a few *R* gene subfamilies. PLoS One 11(2): e0148708

Wenig U, Meyer S, Fischer S, Werner D, Lauter A, Melzer M, Hoth S, Weingartern M, Sauer N (2013) Identification of MAIN, a factor involved in genome stability in the meristems of *Arabidopsis thaliana.* The Plant Journal 75:469-483.

Woo Lim C, Hwan Yang S, Hun Shin K, Chul Lee S, Hyon Kim S (2015) The AtLRK10L1.2 *Arabidopsis* ortholog of wheat LRK10 is involved in ABA-mediated signaling and drought resistance. Plant Cell Reports 34(3):447-455.

Yamada K, Yamaguchi K, Shirakawa K et al. (2016) The Arabidopsis CERK1-associated kinase PBL27 connects chitin perception to MAPK activation. Embo Journal 35:2468-2483.

Yin Z, Ke X, Kang Z, Huang L (2016) Apple resistance responses against Valsa mali revealed by transcriptomics analysis. Physiological and Molecular Plant Pathology 93:85-92.

Table 3.1 Summary of plant material and provenances sequenced in the study.

Tree	Phenotype	Origin
72-132	S	ECC[1]: Southern Chinese (old introduction)
72-139	R[a]	ECC: Southern Chinese (old introduction)
72-41.5	S[b]	ECC: Southern Chinese (old introduction)
72-49.5	S	ECC: Southern Chinese (old introduction)
SC1 aka B66	R	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
SC2	S	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
SC3	S	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
SC4	R	ECC: Southern Chinese (newer introduction; Nanking Botanical Garden)
’Paragon’	HS[c]	ECC: *C. sativa* × *C. dentata*
Paragon-1	HS	ECC: (*C. sativa* × *C. dentata*) × *C. mollissima*
B21	R	ECC: (*C. sativa* × C. *dentata*) × C. *mollissima*
B32	HS	ECC: (*C. sativa* × C. dentata) × C. mollissima
‘Schmucki’	R	ECC: (*C. dentata* × *C. mollissima*) × *C. mollissima* (?) [2]
NC1	S	ECC: B16 (Korean origin) × *C. mollissima*
NC2	S	ECC: B16 (Korean origin) × *C. mollissima*
NC3	S	ECC: *C. dentata* × *C. mollissima* (?)[2]
NC4	R	ECC: B16 (Korean origin) × *C. mollissima*
NC5	R	ECC: B16 (Korean origin) × *C. mollissima*
NC6	R	ECC: B16 (Korean origin)×*C. mollissima*
‘Clapper’	S	TACF: (*C. mollissima* × *C. dentata*) × *C. dentata* “BC1”
‘Nanking’	R	ECC: Southern Chinese (old introduction)
‘Mahogany’	R	TACF: Southern Chinese (old introduction)
Roselawn-1	HS	Northern Indiana, *C. dentata* from outside accepted native range
‘Johnson’	HS	Southern Indiana, *C. dentata* from within native range

[1]Empire Chestnut Company; [2]Pedigree uncertain, inferred from nuclear and chloroplast genotypes

[a]R: resistant; [b]S: susceptible; [c]HS: highly susceptible.

Table 3.2 Summary of reads obtained, estimated assembly depth, and observed average depth for 24 chestnut samples used in the study.

Tree	Phenotype[a]	Reads	Expected avg. depth[b]	Observed avg. depth[c]
72-132	S	156225366	19.53	10.61
72-139	R	123132496	15.39	7.90
72-41.5	S	148922336	18.62	9.92
72-49.5	S	93362230	11.67	10.37
SC1 aka B66	R	216457946	27.06	27.24
SC2	S	82951642	10.37	9.08
SC3	S	204823082	25.60	26.12
SC4	R	223050640	27.88	8.28
’Paragon’	HS	203133080	25.39	23.41
Paragon-1	HS	104247082	13.03	24.68
B21	R	208322230	26.04	24.45
B32	HS	106089926	13.26	12.17
‘Schmucki’	R	223050640	27.88	27.51
NC1	S	221849444	27.73	27.55
NC2	S	56064778	7.01	6.30
NC3	S	190607728	23.83	24.11
NC4	R	207949070	25.99	26.08
NC5	R	202140516	25.27	25.65
NC6	R	104910274	13.11	11.45
‘Clapper’	S	137365152	17.17	8.90
‘Nanking’	R	208326796	26.04	13.12
‘Mahogany’	R	121866930	15.23	8.75
Roselawn-1	HS	348980386	43.62	18.43
‘Johnson’	HS	162119920	20.26	17.12

[a]R = resistant, S = susceptible , HS = highly susceptible; [b]Average read depth expected based on total number of bases in reads divided by genoe size; [c]Observed average read depth in individual genome assemblies.

Table 3.3 SNP and indel calling from whole-genome assemblies.

Pseudo- chromosome	Length	Polymorphic sites	Transitions	Transversions	Ts/Tv	<0.005[a]
LGA	110945115	3072340	2059413	771251	2.67	10910
LGB	42807380	1179877	785277	300165	2.62	4581
LGC	54873616	1673141	1111856	428186	2.60	5274
LGD	60839928	1766935	1187422	444199	2.67	2771
LGE	69694169	1896173	1261689	485153	2.60	2577
LGF	30945471	913696	607872	232108	2.62	2800
LGG	52564869	1495276	1000929	378785	2.64	5000
LGH	51215478	1545270	1031672	392602	2.63	1745
LGI	45177646	1235894	821002	316246	2.60	1830
LGJ	48980084	1426417	950150	360192	2.64	4387
LGK	42513908	1224091	810398	315255	2.57	2212
LGL	73038507	931338	621646	236444	2.63	3600

[a] Count of SNPs associated with variation in blight resistance with significance below the permutation-derived p-value cutoff of 0.005

Table 3.4 SNPs in predicted gene sequences, including 5’ and 3’ UTR and introns.

Pseudo- chromosome	Polymorphic sites	Transitions	Transversions	Ts/Tv	< 0.005[a]
LGA	1745144	1025725	360400	2.85	4224
LGB	660029	382984	137042	2.79	1731
LGC	940807	544332	197470	2.76	1558
LGD	1004764	594059	208103	2.85	984
LGE	1058213	610839	221541	2.76	968
LGF	497503	290014	103796	2.79	1140
LGG	848685	497717	176191	2.82	1566
LGH	870836	503563	181167	2.78	535
LGI	698827	400368	146609	2.73	672
LGJ	810680	477665	169093	2.82	1363
LGK	654244	378339	138323	2.74	881
LGL	531965	312501	111078	2.81	2056

[a] Count of SNPs associated with variation in blight resistance with significance below the permutation-derived p-value cutoff of 0.005

Table 3.5 Polymorphisms identified in exons of predicted genes.

Linkage group	Loci	Ts	Tv	Ts/Tv
LGA	736809	476403	139349	3.42
LGB	279473	177898	54017	3.29
LGC	395908	250863	76816	3.27
LGD	440133	283988	84617	3.36
LGE	450553	285046	88017	3.24
LGF	215304	137444	41670	3.30
LGG	365960	236120	70289	3.36
LGH	367289	232835	71220	3.27
LGI	295255	186343	57673	3.23
LGJ	348615	225950	66528	3.40
LGK	278749	176245	55195	3.19
LGL	229606	146994	44658	3.29

Table 3.6. Uniprot-Swissprot alignments and inferred functions for selected predicted genes in genomic regions with concentrations of blight-resistance associated polymorphisms.

Locus[a]	Gene[b]	Homology-inferred function	Uniprot top hit	ID[c]
LGA.a.1	LGA_g1418	Aberrant root formation protein 4	ALF4_ARATH	44.5
LGA.a.2	LGA_g1419	Aberrant root formation protein 4	ALF4_ARATH	48.6
LGA.a.3	LGA_g1420	Trafficking protein particle complex II-specific subunit 120	TR120_ARATH	80.3
LGA.a.4	LGA_g1421	Homeobox protein knotted-1 like	KNAT3_ARATH	66
LGA.b.1	LGA_g2361	Probable lysophospholipase BODYGUARD 3-like	BDG2_ARATH	58.1
LGA.c.1	LGA_g3528	Probably aquaporin, NIP5-1 like	NIP51_ARATH	80.8
LGA.d.1	LGA_g4191	Leaf rust 10 disease resistance locus RLK	LRL14_ARATH	58
LGA.d.2	LGA_g4193	Leaf rust 10 disease resistance locus RLK	LRL14_ARATH	48.5
LGA.d.3	LGA_g4196	Leaf rust 10 disease resistance locus RLK	LRL27_ARATH	31
LGA.e.1	LGA_g8459	LisH domain and HEAT repeat-containing protein	K1468_DANRE	29
LGA.e.2	LGA_g8465	Serine carboxypeptidase	SCP17_ARATH	44
LGB.a.1	LGB_g2158	G-type lectin S-receptor-like serine/threonine protein kinase	LRL14_ARATH	43.8
LGB.a.2	LGB_g2160	TMV resistance protein, TAO1-like	TAO1_ARATH	36.5
LGB.b.1	LGB_g2214	Protein PIN-LIKES 5	PILS5_ARATH	80
LGB.b.2	LGB_g2245	Cytochrome P450 90B1	C90B1_ARATH	79
LGB.c.1	LGB_g3005	DETOXIFICATION-27, MATE-like	DTX26_ARATH	74.5
LGB.c.2	LGB_g3006	DETOXIFICATION-27, MATE-like	DTX27_ARATH	66.4
LGB.c.3	LGB_g3013	DETOXIFICATION-27, MATE-like	DTX27_ARATH	74.2
LGB.d.1	LGB_g4214	Histidine kinase 1-like	AHK1_ARATH	55.3
LGB.d.2	LGB_g4216	Senescence/dehydration-associated protein	ERD7_ARATH	51.4
LGB.d.3	LGB_g4217	Senescence/dehydration-associated protein	ERD7_ARATH	62.2
LGB.d.4	LGB_g4219	Proteasome non-ATPase regulatory subunit 4	PSMD4_ARATH	57
LGB.e.1	LGB_g5040	F-box protein At4g35733	FB311_ARATH	49
LGB.e.2	LGB_g5043	F-box protein At4g35733	FB311_ARATH	33
LGB.e.3	LGB_g5047	Putative F-box protein	FB72_ARATH	31
LGB.e.4	LGB_g5048	Serine/threonine-protein kinase EDR1	EDR1_ARATH	47
LGB.e.5	LGB_g5051	Putative F-box protein At1g65770	FB72_ARATH	31.6
LGB.e.6	LGB_g5057	Putative F-box protein At1g65770	FB72_ARATH	33
LGC.a.1	LGC_g3384	MLP-like protein 328	ML328_ARATH	66
LGC.a.2	LGC_g3385	MLP-like protein 328	ML328_ARATH	62
LGC.a.3	LGC_g3412	E3 ubiquitin-protein ligase KEG	KEG_ARATH	88
LGC.a.4	LGC_g3428	Probable alpha-mannosidase At5g66150	MANA3_ARATH	72
LGD.a.1	LGD_g1162	Protein EARLY FLOWERING 3	ELF3_ARATH	55
LGD.a.2	LGD_g1165	Clathrin interactor EPSIN 1	EPN1_ARATH	58
LGD.a.3	LGD_g1179	Cationic peroxidase 1	PER1_ARAHY	76

Table 3.6 Continued

Locus[a]	Gene[b]	Homology-inferred function	Uniprot top hit	ID
LGE.a.2	LGE_g7935	Not identifiable from database	na[d]	na
LGE.a.3	LGE_g7940	GDSL esterase/lipase 2	GLIP2_ARATH	56
LGF.a.1	LGF_g1803	Periodic tryptophan protein / WD40 domain protein	PWP2_SCHPO	44
LGF.a.2	LGF_g1804	Periodic tryptophan protein 2 / WD40 domain protein	PWP2_YEAST	59
LGF.b.1	LGF_g3411	Serine/threonine protein phosphatase	PPP7L_ARATH	62
LGF.b.2	LGF_g3412	F-box protein	SKI23_ARATH	31
LGG.a.1	LGG_g2307	Putative ABC transporter B family member 8	AB18B_ARATH	71
LGG.b.1	LGG_g3657	Nicotinamidase 1	NIC1_ARATH	83
LGG.c.2	LGG_g4295	Probable carboxylesterase 5	CXE5_ARATH	45
LGG.d.1	LGG_g6194	Serine/threonine protein kinase	PBL27_ARATH	80
LGG.d.2	LGG_g6195	Serine/threonine protein kinase	PBL27_ARATH	81
LGG.d.3	LGG_g6252	Ethylene-responsive transcription factor	ERF92_ARATH	67
LGG.d.4	LGG_g6270	Ethylene-responsive transcription factor	ERF98_ARATH	77
LGJ.a.1	LGJ_g238	Major pollen allergen / pathogenesis-related protein	PRU1_PRUAR	53
LGJ.a.2	LGJ_g239	Major pollen allergen / pathogenesis-related protein	BEV1D_BETPN	39
LGJ.a.3	LGJ_g240	Cystolic carboxypeptiase 1 (poor alignment)	na	na
LGJ.a.4	LGJ_g243	WD repeat-containing protein 43	WDR43_MOUSE	24
LGK.a.1	LGK_g2007	E3 ubiquitin-protein ligase MIEL1	MIEL1_ARATH	79
LGL.a.2	LGL_g4221	Serine/threonine protein phosphatase MAIN-like	PPP7L_ARATH	58
LGL.a.3	LGL_g4222	Serine/threonine protein phosphatase MAIN-like	PPP7L_ARATH	58
LGL.a.4	LGL_g4223	Armadillo repeat-containing kinesin-like protein	na	57
LGL.a.5	LGL_g4224	Not identifiable from database	na	na
LGL.b.1	LGL_g6433	Creatine kinase U-type	U76E6_ARATH	38
LGL.b.2	LGL_g6436	Probable serine/threonine protein kinase PBL4	PBL4_ARATH	33
LGL.b.3	LGL_g6438	Not identifiable from database	na	na
LGL.c.1	LGL_g6970	Probable disease resistance protein At5g66910	DRL43_ARATH	35
LGL.c.2	LGL_g6971	Probable disease resistance protein At5g66900	DRL42_ARATH	36
LGL.c.3	LGL_g6975	Probable disease resistance protein At5g66900	DRL42_ARATH	37
LGL.c.4	LGL_g6979	Probable disease resistance protein At5g66900	DRL42_ARATH	49
LGL.c.5	LGL_g6992	Probable disease resistance protein At5g66900	DRL42_ARATH	39
LGL.c.6	LGL_g6993	Probable disease resistance protein At5g66900	DRL42_ARATH	29
LGL.c.7	LGL_g6996	Phospholipase D alpha 1	PLDA1_TOBAC	58

Table 3.6 Continued

Locus[a]	Gene[b]	Homology-inferred function	Uniprot top hit	ID
LGL.c.8	LGL_g6997	Probable disease resistance protein At5g66900	DRL42_ARATH	38
LGL.d.1	LGL_g8955	Retrovirus-related Pol polyprotein, transposon TNT 1-94	POLX_TOBAC	54
LGL.d.2	LGL_g8976	Eukaryotic peptide chain release factor subunit 1-3	ERF1Z_ARATH	88
LGL.d.3	LGL_g8978	Eukaryotic peptide chain release factor subunit 1-3	ERF1Z_ARATH	90
LGL.d.4	LGL_g8981	Eukaryotic peptide chain release factor subunit 1-3	ERF1Z_ARATH	90

[a] Regions identified with large numbers of blight-associated SNPs; [b] Gene number designated by AUGUSTUS gene prediction software; [c] Percent identity of amino acids in alignment to the Uniprot-SwissProt database; [d] na indicates that no alignment was found for this sequence.

Table 3.7 Alignments from the nr database for selected predicted genes in genomic regions with concentrations of blight-resistance associated polymorphisms.

Locus	Gene	Identifier[a]	nr top hit[b]	species[c]	ID[d]
LGA.a.1	LGA_g1418	ALF4	XP_018843788.1	*Juglans regia*	68
LGA.a.2	LGA_g1419	ALF4	EOY29388.1	*Theobroma cacao*	64
LGA.a.3	LGA_g1420	TR120	XP_018843784.1	*Juglans regia*	88
LGA.a.4	LGA_g1421	KNAT3	XP_011458839.1	*Fragaria vesca*	81
LGA.b.1	LGA_g2361	BDG2	XP_018814184.1	*Juglans regia*	87
LGA.c.1	LGA_g3528	NIP51	XP_018833777.1	*Juglans regia*	90
LGA.d.1	LGA_g4191	LRL14	XP_018849594.1	*Juglans regia*	76
LGA.d.2	LGA_g4193	LRL14	XP_018816649.1	*Juglans regia*	78
LGA.d.3	LGA_g4196	LRL27	XP_018816648.1	*Juglans regia*	64
LGA.e.1	LGA_g8459	K1468	XP_018849592.1	*Juglans regia*	84
LGA.e.2	LGA_g8465	SCP17	XP_018855843	*Juglans regia*	80
LGB.a.1	LGB_g2158	LRL14	XP_007140615.1	*Phaseolus vulgaris*	61
LGB.a.2	LGB_g2160	TAO1	XP_016540282.1	*Capsicum annuum*	66
LGB.b.1	LGB_g2214	PILS5	XP_018856104.1	*Juglans regia*	92
LGB.b.2	LGB_g2245	C90B1	XP_018826320	*Juglans regia*	90
LGB.c.1	LGB_g3005	DTX26	XP_019074070.1	*Vitis vinifera*	85
LGB.c.2	LGB_g3006	DTX27	XP_008383420.1	*Malus domestica*	85
LGB.c.3	LGB_g3013	DTX27	XP_019074070.1	*Vitis vinifera*	85
LGB.d.1	LGB_g4214	AHK1	XP_018829637.1	*Juglans regia*	88
LGB.d.2	LGB_g4216	ERD7	KDO71260.1	*Citrus sinensis*	86
LGB.d.3	LGB_g4217	ERD7	XP_018809719.1	*Juglans regia*	81
LGB.d.4	LGB_g4219	PSMD4	KYP55024.1	*Cajanus cajan*	78
LGB.e.1	LGB_g5040	FB311	XP_016187763.1	*Arachis ipaensis*	53
LGB.e.2	LGB_g5043	FB311	XP_016187763.1	*Arachis ipaensis*	54
LGB.e.3	LGB_g5047	FB72	GAU34246.1	*Trifolium subterraneum*	55
LGB.e.4	LGB_g5048	EDR1	XP_018825653.1	*Juglans regia*	78
LGB.e.5	LGB_g5051	FB72	XP_016187763.1	*Arachis ipaensis*	62
LGB.e.6	LGB_g5057	FB72	XP_006574655.1	*Glycine max*	63
LGC.a.1	LGC_g3384	ML328	XP_018824604.1	*Juglans regia*	83
LGC.a.2	LGC_g3385	ML328	XP_018824604.1	*Juglans regia*	82
LGC.a.3	LGC_g3412	KEG	KRH62821.1	*Glycine max*	90
LGC.a.4	LGC_g3428	MANA3	XP_018859970.1	*Juglans regia*	40
LGD.a.1	LGD_g1162	ELF3	XP_018834496.1	*Juglans regia*	72
LGD.b.2	LGD_g1165	EPN1	XP_018834513.1	*Juglans regia*	78
LGD.b.3	LGD_g1179	PER1	XP_018844618.1	*Juglans regia*	73
LGD.b.4	LGD_g1185	DTX27	XP_002306557.2	*Populus trichocarpa*	59
LGD.b.1	LGD_g2262	PME61	XP_018805211.1	*Juglans regia*	64
LGD.b.2	LGD_g2276	CRK3	XP_018805237.1	*Juglans regia*	53
LGD.b.3	LGD_g2282	CRK3	XP_018805237.1	*Juglans regia*	88
LGE.a.1	LGE_g7934	CXE15	OJD16848.1	*Emergomyces pasteuriana*	37
LGE.a.2	LGE_g7935	unknown	KDP20420.1	*Jatropha curcas*	34
LGE.a.3	LGE_g7940	GLIP2	XP_010671940.1	*Beta vulgaris*	40
LGF.a.1	LGF_g1803	PWP2	XP_018823554.1	*Juglans regia*	81
LGF.a.2	LGF_g1804	PWP2	GAV65986.1	*Cephalotus follicularis*	89
LGF.b.1	LGF_g3411	PPP7L	XP_018846367.1	*Juglans regia*	36

Table 3.7 continued

Locus	Gene	Identifier[a]	nr top hit[b]	species[c]	ID[d]
LGF.b.2	LGF_g3412	SKI23	XP_018812063.1	Juglans regia	79
LGG.c.1	LGG_g4294	CXE12	XP_004294838.1	*Fragaria vesca*	88
LGG.c.2	LGG_g4295	CXE5	XP_018835580.1	*Juglans regia*	80
LGG.d.1	LGG_g6194	PBL27	XP_018826046.1	*Juglans regia*	91
LGG.d.2	LGG_g6195	PBL27	XP_018840619.1	*Juglans regia*	90
LGG.d.3	LGG_g6252	ERF92	XP_018842116.1	*Juglans regia*	84
LGG.d.4	LGG_g6269	ERF98	XP_003555512.1	*Glycine max*	66
LGG.d.5	LGG_g6270	ERF98	XP_010039174.1	*Eucalyptus grandis*	61
LGJ.a.1	LGJ_g238	PRU1	XP_010039174.1	*Eucalyptus grandis*	61
LGJ.a.2	LGJ_g239	BEV1D	XP_018812583.1	*Juglans regia*	88
LGJ.a.3	LGJ_g240	unknown	XP_018812580.1	*Juglans regia*	70
LGJ.a.4	LGJ_g243	WDR43	na	na	na
LGK.a.1	LGK_g2007	MIEL1	XP_007223130.1	*Prunus persica*	81
LGL.a.2	LGL_g4221	PPP7L	KDO57343.1	*Citrus sinensis*	60
LGL.a.3	LGL_g4222	PPP7L	XP_007204509.1	*Prunus persica*	25
LGL.a.4	LGL_g4223	unknown	XP_014499425.1	*Vigna radiata*	34
LGL.a.5	LGL_g4224	unknown	na	na	na
LGL.b.1	LGL_g6433	U76E6	XP_007200122.1	*Prunus persica*	50
LGL.b.2	LGL_g6436	PBL4	na	na	na
LGL.b.3	LGL_g6438	unknown	na	na	na
LGL.c.1	LGL_g6970	DRL43	XP_018806549.1	*Juglans regia*	52
LGL.c.2	LGL_g6971	DRL42	XP_018806549.1	*Juglans regia*	53
LGL.c.3	LGL_g6975	DRL42	XP_018806549.1	*Juglans regia*	56
LGL.c.4	LGL_g6979	DRL42	XP_018806549.1	*Juglans regia*	75
LGL.c.5	LGL_g6992	DRL42	XP_018806549.1	*Juglans regia*	58
LGL.c.6	LGL_g6993	DRL42	XP_018806549.1	*Juglans regia*	40
LGL.c.7	LGL_g6996	PLDA1	CBI36315.3	*Vitis vinifera*	56
LGL.c.8	LGL_g6997	DRL42	XP_018806549.1	*Juglans regia*	57
LGL.d.1	LGL_g8955	POLX	OMO59210.1	*Corchorus capsularis*	55
LGL.d.2	LGL_g8976	ERF1Z	XP_010264350.1	*Nelumbo nucifera*	93
LGL.d.3	LGL_g8978	ERF1Z	XP_018834951.1	*Juglans regia*	96
LGL.d.4	LGL_g8981	ERF1Z	OMO99264.1	*Corchorus olitorius*	93

[a]Identifiers based on homology to sequences in the Uniprot database; [b]top hit from the non-reduntant (nr) GenBank protein database; [c]species of the top hit; [d]percent identity.

Table 3.8 Total counts of polymorphisms and statistically associated polymorphisms in predicted genes in genomic regions with concentrations of blight-resistance associated polymorphisms.

Locus	Code[a]	Start Coordinate[b]	N (gene)[c]	N (exon)[d]	N (nsyn)[e]	Nassoc (gene)[f]	Nassoc (exon)[g]	Nassoc (nsyn)[h]
LGA.a.1	ALF4	11757760	151	26	19	27	3	2
LGA.a.2	ALF4	11769164	88	0	0	3	0	0
LGA.a.3	TR120	11777361	150	41	14	46	8	4
LGA.a.4	KNAT3	11816491	119	0	0	2	0	0
LGA.b.1	BDG2	19404908	167	6	1	6	0	0
LGA.c.1	NIP51	28672262	493	10	8	5	1	0
LGA.d.1	LRL14	33767628	66	2	0	0	0	0
LGA.d.2	LRL14	33779068	109	4	0	0	0	0
LGA.d.3	LRL27	33804130	57	45	18	6	4	3
LGA.e.1	K1468	66550944	381	128	52	61	19	7
LGA.e.2	SCP17	66617784	123	5	1	9	1	0
LGB.a.1	LRL14	16807295	160	43	25	6	3	1
LGB.a.2	TAO1	16821577	437	208	105	49	16	9
LGB.b.1	PILS5	17298202	97	24	9	0	0	0
LGB.b.2	C90B1	17579356	26	6	2	0	1	1
LGB.c.1	DTX26	23530176	118	5	1	12	0	0
LGB.c.2	DTX27	23538448	19	13	6	1	2	1
LGB.c.3	DTX27	23579417	50	4	2	7	1	0
LGB.d.1	AHK1	33201853	92	32	21	8	1	0
LGB.d.2	ERD7	33231138	50	11	6	8	1	1
LGB.d.3	ERD7	33235012	48	18	3	6	2	0
LGB.d.4	PSMD4	33270656	149	0	0	0	0	0
LGB.e.1	FB311	40299534	135	15	2	0	0	0
LGB.e.2	FB311	40340494	56	26	18	0	1	1
LGB.e.3	FB72	40359334	38	22	1	0	4	0
LGB.e.4	EDR1	40366704	279	68	24	2	2	1
LGB.e.5	FB72	40407624	140	95	8	0	1	0
LGB.e.6	FB72	40446194	68	24	0	0	2	0
LGC.a.1	ML328	27163800	72	17	6	0	0	0
LGC.a.2	ML328	27180063	60	5	0	0	0	0
LGC.a.3	KEG	27396643	5	2	0	0	0	0
LGC.a.4	MANA3	27541167	336	119	37	0	1	0
LGD.a.1	ELF3	8760990	131	35	6	0	0	0
LGD.b.2	EPN1	8780477	150	24	8	0	3	1
LGD.b.3	PER1	8945783	100	12	6	0	2	1
LGD.b.4	DTX27	8984387	61	11	2	0	1	0
LGD.b.1	PME61	17548804	18	9	3	0	0	0

Table 3.8 continued

Locus	Code[a]	Start Coordinate[b]	N (gene)[c]	N (exon)[d]	N (nsyn)[e]	Nassoc (gene)[f]	Nassoc (exon)[g]	Nassoc (nsyn)[h]
LGD.b.2	CRK3	17625144	73	31	8	4	3	1
LGD.b.3	CRK3	17672154	48	0	0	0	0	0
LGE.a.1	CXE15	63363753	32	26	4	6	10	1
LGE.a.2	unknown	63367853	94	16	2	17	9	1
LGE.a.3	GLIP2	63403593	65	0	0	0	0	0
LGF.a.1	PWP2	14762521	57	20	0	0	1	0
LGF.a.2	PWP2	14766961	35	11	0	0	1	0
LGF.b.1	PPP7L	27788524	255	158	65	35	23	5
LGF.b.2	SKI23	27796920	83	0	0	0	0	0
LGG.a.1	AB18B	18414651	124	33	9	0	1	1
LGG.b.1	NIC1	28701241	157	15	10	0	0	0
LGG.c.1	CXE12	33600597	14	8	0	0	1	0
LGG.c.2	CXE5	33604467	301	7	0	1	0	0
LGG.d.1	PBL27	47924998	186	1	1	0	0	0
LGG.d.2	PBL27	47937538	12	2	0	0	0	0
LGG.d.3	ERF92	48342359	6	1	1	0	0	0
LGG.d.4	ERF98	48449428	9	4	0	0	0	0
LGG.d.5	ERF98	48455315	168	4	3	0	0	0
LGJ.a.1	PRU1	1892839	33	0	0	4	0	0
LGJ.a.2	BEV1D	1894532	13	0	0	4	0	0
LGJ.a.3	unknown	1895988	125	50	21	25	4	3
LGJ.a.4	WDR43	1918982	126	17	7	9	0	0
LGK.a.1	MIEL1	16263752	73	0	0	0	0	0
LGL.a.2	PPP7L	33571090	228	137	30	28	16	4
LGL.a.3	PPP7L	33578202	103	9	5	62	9	5
LGL.a.4	unknown	33591325	179	40	17	137	39	17
LGL.a.5	unknown	33598775	71	51	12	64	45	12
LGL.b.1	U76E6	51602920	37	8	0	17	6	0
LGL.b.2	PBL4	51630590	109	16	1	9	1	0
LGL.b.3	unknown	51641220	33	3	0	3	0	0
LGL.c.1	DRL43	55667083	264	109	42	0	0	0
LGL.c.2	DRL42	55678755	200	50	21	0	0	0
LGL.c.3	DRL42	55711942	65	40	20	0	1	0
LGL.c.4	DRL42	55743715	12	9	3	0	1	0
LGL.c.5	DRL42	55852472	178	26	17	0	2	0
LGL.c.6	DRL42	55862625	24	9	4	0	0	0
LGL.c.7	PLDA1	55876970	238	63	13	4	3	2
LGL.c.8	DRL42	55884860	56	20	8	0	3	2

Table 3.8 continued

Locus	Code[a]	Start Coordinate[b]	N (gene)[c]	N (exon)[d]	N (nsyn)[e]	Nassoc (gene)[f]	Nassoc (exon)[g]	Nassoc (nsyn)[h]
LGL.d.1	POLX	71632437	292	0	0	0	0	0
LGL.d.2	ERF1Z	71832559	122	18	15	0	3	2
LGL.d.3	ERF1Z	71842412	105	19	16	5	9	6
LGL.d.4	ERF1Z	71869001	60	0	0	0	0	0

[a] Identifiers based on homology to sequences in the Uniprot database; [b] start coordinates on pseudochromosome sequences; [c] number of SNPs in the predicted gene; [d] number of SNPs in the predicted exon; [e] number of predicted nonsynonymous SNPs in the gene, [f] number of SNPs within the gene with statistical association p <0.001, [g] number of SNPs within exons with statistical association p <0.005; [h] number of nonsynonymous SNPs with statistical association p<0.005.

Table 3.9 Transcriptome alignments (*Cm* and *Cc*) for selected predicted proteins in genomic regions with concentrations of blight-resistance associated polymorphisms.

Locus	Gene[a]	*Cm*[b]	*Cm* %ID	*Cm* top contig	DE[c]	*Cc*[d]	*Cc* % ID	*Cc* top contig
LGA.a.1	ALF4	2	100	CCall_contig31852_v2	-	1	96.4	isotig06361
LGA.a.2	ALF4	0	-	-	-	2	100	isotig01518*
LGA.a.3	TR120	6	100	Ccall_contig47536_v2	-	1	92.3	isotig07234
LGA.a.4	KNAT3	2	81.4	CCall_contig26097_v2	-	0	93.1	isotig01304*
LGA.b.1	BDG2	2	100	Ccall_contig44711_v2	-	0	-	-
LGA.c.1	NIP51	1	100	Ccall_contig18565_v2	down	1	94.2	isotig03352
LGA.d.1	LRL14	0	-	-	-	0	-	-
LGA.d.2	LRL14	4	93.5	Ccall_contig28610_v2	-	3	88.9	isotig00386
LGA.d.3	LRL27	1	91.2	Ccall_contig30127_v2	-	0	-	-
LGA.e.1	K1468	7	100	Ccall_contig42355_v2	up	1	90.4	isotig07188
LGA.e.2	SCP17	1	100	CCall_contig25043_v2	up	0	-	-
LGB.a.1	LRL14	0	-	-	-	0	-	-
LGB.a.2	TAO1	0	-	-	-	0	-	-
LGB.b.1	PILS5	2	98.2	CCall_contig38566_v2	up	2	95.5	isotig02647
LGB.b.2	C90B1	0	-	-	-	0	-	-
LGB.c.1	DTX26	0	-	-	-	0	-	-
LGB.c.2	DTX27	0	-	-	-	0	-	-
LGB.c.3	DTX27	0	-	-	-	0	-	-
LGB.d.1	AHK1	2	99.5	Ccall_contig34660_v2	-	0	-	-
LGB.d.2	ERD7	0	-	-	-	1	100	isotig07342
LGB.d.3	ERD7	2	100	Ccall_contig43999_v2	-	2	93.2	isotig03485
LGB.d.4	PSMD4	2	100	CCall_contig27137_v2	-	0	-	-
LGB.e.1	FB311	1	94.2	CCall_contig15713_v2	-	0	-	-
LGB.e.2	FB311	1	78.9	CCall_contig18139_v2	-	1	83.8	isotig04725
LGB.e.3	FB72	1	96.5	CCall_contig8323_v2	-	0	-	-
LGB.e.4	EDR1	0	-	-	-	0	-	-
LGB.e.5	FB72	2	100	Ccall_contig47193_v2	-	0	-	-
LGB.e.6	FB72	2	92.7	Ccall_contig46838_v2	-	0	-	-
LGC.a.1	ML328	1	95.9	CCall_contig22158_v2	-	4	100	isotig04418
LGC.a.2	ML328	1	100	CCall_contig17986_v2	-	0	-	-
LGC.a.3	KEG	2	99	CCall_contig31162_v2	-	0	-	-
LGC.a.4	MANA3	1	90.6	CCall_contig10436_v2	-	0	-	-
LGD.a.1	ELF3	4	100	CCall_contig31586_v2	-	0	-	-
LGD.b.2	EPN1	2	96.9	CCall_contig2506_v2	-	0	-	-
LGD.b.3	PER1	2	100	CCall_contig8733_v2	-	0	-	-
LGD.b.4	DTX27	1	100	CCall_contig29371_v2	-	0	-	-
LGD.b.1	PME61	1	100	Ccall_contig34356_v2	-	0	-	-
LGD.b.2	CRK3	0	-	-	-	0	-	-

Table 3.9 continued

Locus	Gene[a]	*Cm*[b]	*Cm* % ID	*Cm* top contig	DE[c]	*Cc*[d]	*Cc* % ID	*Cc* top contig
LGE.a.1	CXE15	0	-	-	-	0	-	-
LGE.a.2	na[e]	0	-	-	-	0	-	-
LGE.a.3	GLIP2	1	100	Ccall_contig3408_v2	-	0	-	-
LGF.a.1	PWP2	0	-	-	-	0	-	-
LGF.a.2	PWP2	2	100	CCall_contig41440_v2	-	0	-	-
LGF.b.1	PPP7L	0	-	-	-	0	-	-
LGF.b.2	SKI23	2	99.3	Ccall_contig46622_v2	-	0	-	-
LGG.a.1	AB18B	1	100	CCall_contig19534_v2	-	0	-	-
LGG.b.1	NIC1	2	98	CCall_contig40455_v2	up	0	-	-
LGG.c.1	CXE12	1	98.2	CCall_contig5720_v2	-	0	-	-
LGG.c.2	CXE5	1	98.7	CCall_contig40179_v2	up	0	-	-
LGG.d.1	PBL27	0	-	-	-	0	-	-
LGG.d.2	PBL27	1	100	Ccall_contig43668_v2	-	1	99.7	isotig02337
LGG.d.3	ERF92	1	86.9	Ccall_contig34756_v2	-	1	86	isotig05666*
LGG.d.4	ERF98	0	-	-	-	0	-	-
LGG.d.5	ERF98	0	-	-	-	0	-	-
LGJ.a.1	PRU1	1	100	Ccall_contig37079_v2	-	1	99.4	isotig05884
LGJ.a.2	BEV1D	0	-	-	-	0	-	-
LGJ.a.3	na	0	-	-	-	0	-	-
LGJ.a.4	WDR43	2	97.7	Ccall_contig35765_v2	-	0	-	-
LGK.a.1	MIEL1	1	97.8	CCall_contig9398_v2	up	1	66.8	isotig04345
LGL.a.2	PPP7L	0	-	-	-	0	-	-
LGL.a.3	PPP7L	0	-	-	-	0	-	-
LGL.a.4	na	0	-	-	-	0	-	-
LGL.a.5	na	0	-	-	-	0	-	-
LGL.b.1	U76E6	0	-	-	-	0	-	-
LGL.b.2	PBL4	1	81.8	Ccall_contig21142_v2	-	0	-	-
LGL.b.3	na	0	-	-	-	0	-	-
LGL.c.1	DRL43	1	63.7	CCall_contig27259_v2	-	1	86	isotig06048
LGL.c.2	DRL42	2	100	CCall_contig43338_v2	-	1	92.8	isotig01936*
LGL.c.3	DRL42	0	0	0	-	0	-	-
LGL.c.4	DRL42	1	88.4	CCall_contig5320_v2	-	0	-	-
LGL.c.5	DRL42	2	97.3	CCall_contig2794_v2	-	2	98.9	isotig06133
LGL.c.6	DRL42	1	99.1	CCall_contig34307_v2	-	0	-	-
LGL.c.7	PLDA1	1	63.2	CCall_contig30890_v2	-	0	-	-
LGL.c.8	DRL42	0	0	0	-	1	89.2	isotig06565
LGL.d.1	POLX	0	-	-	-	0	-	-
LGL.d.2	ERF1Z	0	-	-	-	1	92.9	isotig00289
LGL.d.3	ERF1Z	1	89.9	CCall_contig8884_v2	-	0	-	-

Table 3.9 continued

Locus	Gene[a]	*Cm*[b]	*Cm* % ID	*Cm* top contig	DE[c]	*Cc*[d]	*Cc* % ID	*Cc* top contig
LGL.d.4	ERF1Z	2	90	CCall_contig44799_v2	-	0	-	-

[a]Identifiers based on homology to Uniprot; [b]number of alignments to publicly available *Castanea mollissima* (*Cm*) (Barakat et al. 2012) transcriptome; ; [c]alignments to contigs that were differentially expressed in healthy versus blight-inoculated *Cm* stem tissue (Barakat et al. 2012); [d]alignments to *Castanea crenata* (*Cc*) (Serrazina et al. 2015) cDNA contigs, with name and percent identity of the best cDNA contig alignment from each species; [3]No putative gene function assigned due to poor alignment to UniProt database. *Differentially expressed in response to *Phytophthora* root rot (Serrazina et al. 2015)

Table 3.10 Transcriptome alignments (*Cd* and *Cs*) for predicted genes in genomic regions with concentrations of blight-resistance associated polymorphisms.

Locus	Gene[a]	*Cd*[b]	*Cd* top %	*Cd* top contig	DE[c]	*Cs*[d]	*Cs* top %	*Cs* top contig
LGA.a.1	ALF4	1	92.6	AC454_contig26902_v2	-	0	-	-
LGA.a.2	ALF4	0	-	-	-	0	-	-
LGA.a.3	TR120	1	80.3	AC454_contig5201_v2	-	2	94.2	isotig04996
LGA.a.4	KNAT3	3	100	AC454_contig33557_v2	-	2	97.9	isotig05734
LGA.b.1	BDG2	3	98.4	AC454_contig14318_v2	-	0	-	-
LGA.c.1	NIP51	2	94.2	AC454_contig31353_v2	-	1	93.8	isotig03295
LGA.d.1	LRL14	1	95.9	AC454_contig32412_v2	up	1	94.6	isotig02376
LGA.d.2	LRL14	3	100	AC454_contig14818_v2	-	3	95	isotig02356
LGA.d.3	LRL27	1	72.5	AC454_contig22138_v2	-	0	-	-
LGA.e.1	K1468	1	86.2	AC454_contig8158_v2	-	0	-	-
LGA.e.2	SCP17	0	-	-	-	0	-	-
LGB.a.1	LRL14	0	-	-	-	0	-	-
LGB.a.2	TAO1	0	-	-	-	0	-	-
LGB.b.1	PILS5	1	94.1	AC454_contig7351_v2	-	2	93.6	isotig02311
LGB.b.2	C90B1				-			
LGB.c.1	DTX26	0	-	-	-	0	-	-
LGB.c.2	DTX27	0	-	-	-	0	-	-
LGB.c.3	DTX27	0	-	-	-	0	-	-
LGB.d.1	AHK1	5	100	AC454_contig33369_v2	-	0	-	-
LGB.d.2	ERD7	1	65.4	AC454_contig18784_v2	-	1	85.9	isotig06934
LGB.d.3	ERD7	2	78.9	AC454_contig23832_v2	-	1	88.3	isotig04394
LGB.d.4	PSMD4	1	74.6	AC454_contig33859_v2	up	0	-	-
LGB.e.1	FB311	0	-	-	-	0	-	-
LGB.e.2	FB311	0	-	-	-	0	-	-
LGB.e.3	FB72	0	-	-	-	0	-	-
LGB.e.4	EDR1	1	100	AC454_contig1831_v2	-	0	-	-
LGB.e.5	FB72	1	78.3	AC454_contig14527_v2	-	1	84.7	isotig05782
LGB.e.6	FB72	2	98.8	AC454_contig16808_v2	-	1	97.7	isotig07703
LGC.a.1	ML328	0	-	-	-	3	100	isotig01644
LGC.a.2	ML328	0	-	-	-	0	-	-
LGC.a.3	KEG	0	-	-	-	0	-	-
LGC.a.4	MANA3	0	-	-	-	0	-	-
LGD.a.1	ELF3	1	100	AC454_contig941_v2	up	0	-	-
LGD.b.2	EPN1	2	98.3	AC454_contig23664_v2	-	2	99.1	isotig05257
LGD.b.3	PER1	1	100	AC454_contig12667_v2	-	0	-	-
LGD.b.4	DTX27	0	-	-	-	0	-	-
LGD.b.1	PME61	0	-	-	-	0	-	-

Table 3.10 continued

Locus	Gene[a]	*Cd*[b]	*Cd* top %	*Cd* top contig	DEG[c]	*Cs*[d]	*Cs* top %	*Cs* top contig
LGD.b.2	CRK3	0	-	-	-	0	-	-
LGD.b.3	CRK3	1	100	AC454_contig20090_v2	-	1	98.8	isotig04723
LGE.a.1	CXE15	0	-	-	-	0	-	-
LGE.a.2	na	0	-	-	-	0	-	-
LGE.a.3	GLIP2	1	100	AC454_contig20090_v2	-	0	-	-
LGF.a.1	PWP2	1	100	AC454_contig15760_v2	-	0	-	-
LGF.a.2	PWP2	0	-	-	-	0	-	-
LGF.b.1	PPP7L	0	-	-	-	0	-	-
LGF.b.2	SKI23	3	98.8	AC454_contig16593_v2	-	0	-	-
LGG.a.1	AB18B	0	-	-	-	0	-	-
LGG.b.1	NIC1	1	85.8	AC454_contig5108_v2	-	0	-	-
LGG.c.1	CXE12	0	-	-	-	0	-	-
LGG.c.2	CXE5	3	99.3	AC454_contig34315_v2	-	0	-	-
LGG.d.1	PBL27	1	71.4	AC454_contig13148_v2	-	0	-	-
LGG.d.2	PBL27	1	98.8	AC454_contig11118_v2	-	0	-	-
LGG.d.3	ERF92	4	100	AC454_contig32650_v2	up	0	-	-
LGG.d.4	ERF98	0	-	-	-	0	-	-
LGG.d.5	ERF98	0	-	-	-	0	-	-
LGJ.a.1	PRU1	1	99.4	AC454_contig17310_v2	-	1	99.4	isotig05368
LGJ.a.2	BEV1D	0	-	-	-	0	-	-
LGJ.a.3	na	0	-	-	-	0	-	-
LGJ.a.4	WDR43	1	96	AC454_contig18841_v2	-	0	-	-
LGK.a.1	MIEL1	0	-	-	-	0	-	-
LGL.a.2	PPP7L	0	-	-	-	0	-	-
LGL.a.3	PPP7L	0	-	-	-	0	-	-
LGL.a.4	na	0	-	-	-	0	-	-
LGL.a.5	na	0	-	-	-	0	-	-
LGL.b.1	U76E6	0	-	-	-	0	-	-
LGL.b.2	PBL4	0	-	-	-	0	-	-
LGL.b.3	na	1	97.6	AC454_contig26989_v2	-	0	-	-
LGL.c.1	DRL43	1	92	AC454_contig2808_v2	-	1	86	isotig06048
LGL.c.2	DRL42	3	96.5	AC454_contig31156_v2	up	1	92.8	isotig01936*
LGL.c.3	DRL42	1	98.1	AC454_contig22166_v2	-	0	-	-
LGL.c.4	DRL42	0	0	0	-	0	-	-
LGL.c.5	DRL42	0	0	0	-	2	97.6	isotig02122*
LGL.c.6	DRL42	1	79.3	AC454_contig31352_v2	-	0	-	-
LGL.c.7	PLDA1	0	0	0	-	0	-	-
LGL.c.8	DRL42	1	94	AC454_contig15280_v2	-	1	71.2	isotig05945*
LGL.d.1	POLX	1	50	AC454_contig9952_v2	up	0	-	-

Table 3.10 continued

Locus	Code[a]	*Cd*[b]	*Cd* top %	*Cd* top contig	DE[c]	*Cs*[d]	*Cs* top %	*Cs* top contig
LGL.d.2	ERF1Z	1	93.6	AC454_contig10459_v2	-	1	96.5	isotig02801
LGL.d.3	ERF1Z	1	69	AC454_contig19653_v2	-	0	-	-
LGL.d.4	ERF1Z	0	-	-	-	0	-	-

[a]Identifiers based on homology to Uniprot [b]alignments to publicly available *Castanea dentata* (*Cd*) (Barakat et al. 2012) cDNA contigs; [c] contigs that were differentially expressed in healthy versus blight-inoculated *Cd* stem tissue (Barakat et al. 2012); [d]alignments to *Castanea sativa* (*Cs*) (Serrazina et al. 2015) cDNA contigs, with percent identity of the best cDNA contig alignment from each species.

* Differentially expressed in roots of *Cs* inoculated with *Phytophthora cinammomi* (Serrazina et al. 2015).

Table 3.11 Summary of Tajima's D statistic in assemblies of the cbr blight resistance QTL regions (Kubisiak et al. 2013) and genes chosen for concentrations of associated SNPs for each region.

	cbr1: MATE-like gene (LGB.b.3)	cbr1 average	cbr2: NBS-LRR gene	cbr2 average	cbr3: Epoxide hydrolase gene	CBR3 average
Resistant Cm	1.061	1.643	1.052	1.677	1.275	1.714
Susceptible Cm	-0.935	0.901	1.091	1.457	-0.190	1.680
Non-Cm	0.644	1.234	0.602	1.341	0.791	1.641

Table 3.12 Tajima's *D* statistic and π for selected predicted genes in genomic regions with concentrations of blight-resistance associated polymorphisms.

Locus	Code[a]	*D*-Cmr[b]	*D*-Cms[c]	*D*-Cdx[d]	π-Cmr[b]	π-Cms[c]	π-Cdx[d]
LGA.a.1	ALF4	1.985	1.567	-0.138	0.00424	0.00473	0.00535
LGA.a.2	ALF4	0.591	0.915	-0.414	0.00561	0.00500	0.00665
LGA.a.3	TR120	2.376	-1.038	nan	0.00106	nc	0.00131
LGA.a.4	KNAT3	1.835	0.382	0.547	0.00199	nc	0.00207
LGA.b.1	BDG2	0.543	0.904	-0.723	0.00279	0.00498	0.00355
LGA.c.1	NIP51	nan	nan	nan	0.00228	nc	0.00279
LGA.d.1	LRL14	0.854	0.994	1.801	0.00490	0.00510	0.00530
LGA.d.2	LRL14	2.163	0.285	1.298	0.00544	0.00547	0.00459
LGA.d.3	LRL27	2.121	0.554	0.317	0.01619	0.00951	0.01200
LGA.e.1	K1468	-0.306	1.349	0.059	0.00344	0.00539	0.00314
LGA.e.2	SCP17	0.536	1.282	-0.018	nc	nc	nc
LGB.a.1	LRL14	2.214	2.678	0.117	0.01244	0.01266	0.00756
LGB.a.2	TAO1	2.081	2.108	-0.011	0.02313	0.02366	0.01485
LGB.b.1	PILS5	1.366	1.206	-0.483	0.00579	0.00334	0.00806
LGB.b.2	C90B1	0.506	-1.281	-0.186	0.00141	0.00080	0.00392
LGB.c.1	DTX26	1.388	-0.744	-0.979	0.00517	0.00761	0.00464
LGB.c.2	DTX27	0.726	-1.206	-1.234	0.00215	0.00193	0.00163
LGB.c.3	DTX27	0.459	-0.210	-1.145	0.00435	0.00570	0.00378
LGB.d.1	AHK1	-1.054	-0.883	nan	0.00123	0.00221	0.00192
LGB.d.2	ERD7	-0.587	-0.746	-1.081	0.00245	0.00450	0.00345
LGB.d.3	ERD7	0.723	0.603	0.386	0.00551	0.00589	0.00385
LGB.d.4	PSMD4	1.249	-1.148	-0.311	0.00123	0.00123	0.00193
LGB.e.1	FB311	0.627	1.135	1.500	0.00608	0.00593	0.00859
LGB.e.2	FB311	0.724	1.411	1.624	0.00746	0.00490	0.00795
LGB.e.4	EDR1	0.927	0.843	-0.173	0.00709	0.00646	0.01215
LGB.e.5	FB72	2.045	-0.321	-1.187	0.00963	0.02601	0.02721
LGB.e.6	FB72	1.154	0.700	-0.067	0.00783	0.00832	0.01427
LGC.a.1	ML328	1.475	0.281	-0.089	0.00895	0.00950	0.01182
LGC.a.2	ML328	1.584	-0.573	-0.316	0.00568	0.00516	0.00909
LGC.a.3	KEG	0.015	-1.054	nan	0.00011	0.00039	0.00058
LGC.a.4	MANA3	1.348	1.491	0.479	0.01410	0.01263	0.01566
LGD.a.1	ELF3	1.236	1.720	nan	0.00595	0.00169	0.01017
LGD.a.2	EPN1	0.270	1.970	nan	0.00768	0.00259	0.00862
LGD.a.3	PER1	0.015	1.696	nan	0.00826	0.00276	0.00932
LGD.a.4	DTX27	2.201	1.216	nan	0.00457	0.00055	0.00149
LGD.b.1	PME61	1.797	1.163	-1.233	0.00465	0.00443	0.00110
LGD.b.2	CRK3	0.641	-0.055	nan	0.00415	0.00494	0.00093
LGD.b.3	CRK3	0.453	-0.329	nan	0.00149	0.00149	0.00093
LGE.a.1	CXE15	2.106	-0.488	0.030	0.00665	0.00399	0.00028

Table 3.12 continued

Locus	Code[a]	*D*-Cmr[b]	*D*-Cms[c]	*D*-Cdx[d]	π-Cmr[b]	π-Cms[c]	π-Cdx[d]
LGE.a.2	unknown	2.373	-1.234	nan	0.00661	0.00279	0.00007
LGE.a.3	GLIP2	nan	nan	nan	0.00179	0.00179	0.00122
LGF.a.1	PWP2	1.311	1.197	-0.403	0.00244	0.00417	0.00620
LGF.a.2	PWP2	1.671	2.002	-0.286	0.00331	0.00477	0.00711
LGF.b.1	PPP7L	0.641	0.501	-0.843	0.00507	0.01076	0.00813
LGF.b.2	SKI23	1.567	nan	nan	0.00154	0.00154	0.00107
LGG.a.1	AB18B	-1.373	-0.480	-1.110	0.00139	0.00220	0.00577
LGG.b.1	NIC1	0.835	2.183	-0.492	0.00959	0.00711	0.01456
LGG.c.1	CXE12	1.559	0.874	-0.255	0.00162	0.00170	0.00213
LGG.c.2	CXE5	1.288	0.710	0.848	0.00358	0.01185	0.01267
LGG.d.1	PBL27	0.203	-1.480	-0.856	0.00110	0.00411	0.00626
LGG.d.2	PBL27	nan	-1.370	0.176	0.00029	0.00063	0.00222
LGG.d.3	ERF92	1.060	-1.724	-0.673	0.00208	0.00433	0.00510
LGG.d.4	ERF98	nan	-1.780	-1.132	na	0.00056	0.00067
LGG.d.5	ERF98	-1.165	-1.625	-0.933	0.00011	0.00108	0.00067
LGJ.a.1	PRU1	1.791	-0.771	-1.220	0.00209	0.00348	0.00527
LGJ.a.2	BEV1D	nan	-0.970	-0.719	0.00441	0.00223	0.00260
LGJ.a.3	unknown	1.786	1.993	-0.250	0.01166	0.00867	0.01147
LGJ.a.4	WDR43	nan	-0.937	-1.214	nc	nc	nc
LGK.a.1	MIEL1	1.383	nan	nan	0.00057	0.00057	0.00108
LGL.a.2	PPP7L	2.514	-0.638	nan	0.01149	0.00372	0.00742
LGL.a.3	PPP7L	nan	nan	nan	0.00728	0.00262	0.00097
LGL.a.4	unknown	2.891	-0.941	nan	0.01307	0.00244	0.00070
LGL.a.5	unknown	2.922	-0.698	nan	0.01282	0.00254	na
LGL.b.1	U76E6	2.373	-0.578	0.130	0.00506	0.00047	0.00090
LGL.b.2	PBL4	1.749	-1.041	-0.790	nc	nc	nc
LGL.c.1	DRL43	2.539	1.582	1.382	0.01931	0.01722	0.02272
LGL.c.2	DRL42	-0.231	-1.245	0.636	0.00664	0.01043	0.01498
LGL.c.3	DRL42	-0.417	-0.566	0.299	0.00394	0.00407	0.00791
LGL.c.4	DRL42	0.007	-0.434	-1.182	0.00170	0.00104	0.00277
LGL.c.5	DRL42	1.357	-0.693	nan	0.00537	0.00424	0.01193
LGL.c.6	DRL42	2.318	-0.395	1.187	0.00248	0.00211	0.00473
LGL.c.7	PLDA1	1.364	0.341	0.906	0.01015	0.00554	0.01249
LGL.c.8	DRL42	0.575	-0.063	nan	0.00274	0.00141	0.00378
LGL.d.1	POLX	nan	nan	1.024	0.00317	0.00317	0.00430
LGL.d.2	ERF1Z	1.399	2.524	-0.824	0.00469	0.00506	0.01096
LGL.d.3	ERF1Z	2.126	2.337	-0.740	0.00663	0.00542	0.00970
LGL.d.4	ERF1Z	0.899	1.378	-0.635	0.00291	0.00492	0.00710

[a]Identifiers based on homology to Uniprot, [b]Tajima's *D* statistic and nucleotide divergence (π) for highly resistant Chinese chestnuts (Cmr), [c]susceptible Chinese

chestnuts (Cms), and [d]American chestnut / ‘Paragon’ (Cdx). nan = insufficient polymorphism to calculate; nc = not calculated.

Table 3.13 Heterozygosity and interspecific F_{ST} for selected predicted genes in genomic regions with concentrations of blight-resistance associated polymorphisms.

Locus	Code[a]	Het-Hyb[b]	Het-Cmr[b]	Het-Cms[c]	Het-Cd[d]	Fst[e]
LGA.a.1	ALF4	0.300	0.312	0.157	0.184	0.517
LGA.a.2	ALF4	0.431	0.191	0.174	0.168	0.622
LGA.a.3	TR120	0.327	0.374	0.137	0.181	0.585
LGA.a.4	KNAT3	0.375	0.310	0.181	0.157	0.464
LGA.b.1	BDG2	0.352	0.121	0.134	0.108	0.867
LGA.c.1	NIP51	0.430	0.136	0.129	0.200	0.643
LGA.d.1	LRL14	0.303	0.108	0.296	0.315	0.628
LGA.d.2	LRL14	0.257	0.300	0.252	0.284	0.604
LGA.d.3	LRL27	0.427	0.078	0.095	0.074	0.016
LGA.e.1	K1468	0.336	0.077	0.124	0.126	0.740
LGA.e.2	SCP17	0.426	0.314	0.279	0.266	nc
LGB.a.1	LRL14	0.565	0.327	0.240	0.124	0.577
LGB.a.2	TAO1	0.519	0.412	0.338	0.209	0.594
LGB.b.1	PILS5	0.456	0.303	0.179	0.092	0.345
LGB.b.2	C90B1	0.507	0.082	0.111	0.030	0.649
LGB.c.1	DTX26	0.367	0.120	0.215	0.090	0.782
LGB.c.2	DTX27	0.461	0.140	0.178	0.123	0.664
LGB.c.3	DTX27	0.295	0.229	0.253	0.120	0.596
LGB.d.1	AHK1	0.546	0.109	0.208	0.152	0.843
LGB.d.2	ERD7	0.445	0.136	0.210	0.160	0.788
LGB.d.3	ERD7	0.292	0.372	0.278	0.129	0.453
LGB.d.4	PSMD4	0.349	0.302	0.180	0.053	0.550
LGB.e.1	FB311	0.397	0.271	0.242	0.158	0.293
LGB.e.2	FB311	0.461	0.426	0.488	0.331	0.207
LGB.e.3	FB72	0.395	0.325	0.184	0.191	0.350
LGB.e.4	EDR1	0.402	0.222	0.154	0.228	0.395
LGB.e.5	FB72	0.797	0.232	0.174	0.807	0.238
LGB.e.6	FB72	0.479	0.204	0.192	0.212	0.390
LGC.a.1	ML328	0.663	0.435	0.384	0.458	0.213
LGC.a.2	ML328	0.469	0.183	0.194	0.169	0.343
LGC.a.3	KEG	0.508	0.133	0.031	0.267	0.217
LGC.a.4	MANA3	0.513	0.451	0.424	0.382	0.030
LGD.a.1	ELF3	0.522	0.282	0.328	0.041	0.492
LGD.b.2	EPN1	0.431	0.232	0.346	0.062	0.428
LGD.b.3	PER1	0.392	0.253	0.196	0.052	0.451
LGD.b.4	DTX27	0.119	0.628	0.423	0.044	0.210
LGD.b.1	PME61	0.403	0.525	0.271	0.056	0.521
LGD.b.2	CRK3	0.541	0.230	0.216	0.009	0.725
LGD.b.3	CRK3	0.531	0.185	0.245	0.056	0.588

Table 3.13 continued

Locus	Code[a]	Het-Hyb[b]	Het-Cmr[b]	Het-Cms[c]	Het-Cd[d]	Fst[e]
LGE.a.1	CXE15	0.043	0.423	0.191	0.046	0.191
LGE.a.2	unknown	0.000	0.529	0.142	0.037	0.060
LGE.a.3	GLIP2	0.444	0.521	0.557	0.194	0.049
LGF.a.1	PWP2	0.565	0.273	0.149	0.285	0.318
LGF.a.2	PWP2	0.493	0.278	0.181	0.292	0.354
LGF.b.1	PPP7L	0.296	0.141	0.179	0.178	0.802
LGF.b.2	SKI23	0.396	0.204	0.176	0.121	0.521
LGG.a.1	AB18B	0.307	0.067	0.092	0.145	0.626
LGG.b.1	NIC1	0.478	0.206	0.113	0.216	0.252
LGG.c.1	CXE12	0.161	0.286	0.145	0.196	0.119
LGG.c.2	CXE5	0.339	0.256	0.046	0.342	0.269
LGG.d.1	PBL27	0.510	0.036	0.146	0.191	0.863
LGG.d.2	PBL27	0.500	0.028	0.157	0.722	0.761
LGG.d.3	ERF92	0.520	0.074	0.142	0.136	0.851
LGG.d.4	ERF98	0.611	0.000	0.125	0.148	0.939
LGG.d.5	ERF98	0.708	0.019	0.167	0.056	0.964
LGJ.a.1	PRU1	0.561	0.176	0.202	0.260	0.786
LGJ.a.2	BEV1D	0.426	0.239	0.231	0.288	0.731
LGJ.a.3	unknown	0.340	0.344	0.272	0.195	0.800
LGJ.a.4	WDR43	0.525	0.237	0.295	0.218	nc
LGK.a.1	MIEL1	0.539	0.117	0.184	0.333	0.837
LGL.a.2	PPP7L	0.261	0.544	0.170	0.078	0.834
LGL.a.3	PPP7L	0.645	0.226	0.061	0.094	0.291
LGL.b.1	U76E6	0.466	0.461	0.014	0.045	0.477
LGL.b.2	PBL4	0.824	0.224	0.043	0.063	nc
LGL.c.1	DRL43	0.664	0.676	0.585	0.547	0.143
LGL.c.2	DRL42	0.553	0.114	0.159	0.358	0.424
LGL.c.3	DRL42	0.481	0.138	0.194	0.227	0.358
LGL.c.4	DRL42	0.354	0.176	0.177	0.063	0.511
LGL.c.5	DRL42	0.521	0.116	0.154	0.160	0.495
LGL.c.6	DRL42	0.418	0.285	0.151	0.126	0.393
LGL.c.7	PLDA1	0.311	0.356	0.241	0.111	0.300
LGL.c.8	DRL42	0.260	0.396	0.185	0.079	0.375
LGL.d.1	POLX	0.704	0.181	0.130	0.235	0.795
LGL.d.2	ERF1Z	0.564	0.240	0.085	0.209	0.402
LGL.d.3	ERF1Z	0.443	0.462	0.125	0.156	0.218
LGL.d.4	ERF1Z	0.600	0.252	0.079	0.317	0.429

[a] Identifiers based on homology to Uniprot [b] proportion of heterozygous SNPs (Het) calculated among groups for highly resistant Chinese chestnuts (Cmr); [c] susceptible

Chinese chestnuts (Cms); and [d]American chestnut / ‘Paragon’ (Cdx); [e]Interspecific F_{ST}; [f]nc = not calculated.

Table 3.14 Average Tajima's *D* and heterozygosity for predicted genes with homology to known resistance-associated genes in different categories of trees.

Type	Genes surveyed[a]	Total predicted	*D*-Cmr	*D*-Cms	*D*-Cdx	Het-Hyb	Het-Cm	Het-Cdx
LRK10-like	29	61	0.289	0.415	0.513	0.264	0.264	0.249
NBS-LRR	218	545	0.918	0.946	0.326	0.413	0.283	0.224
MATE-like	42	82	0.466	0.541	0.174	0.449	0.187	0.150
Lectin RK	261	519	1.050	0.890	0.345	0.425	0.267	0.189
CytP450	121	312	0.699	0.607	0.077	0.428	0.209	0.176

[a] Only genes found in clusters of 3 or more were included in these calculations.

Table 3.15 Associated SNPs ($p < 0.005$) found upstream of selected blight resistance candidate genes at three kilobase (kb) ranges.

Locus	Identifier	< 5 kb	< 1 kb	< 500 b
LGA.a.1	ALF4	10	1	1
LGA.b.1	BDG2	2	0	0
LGA.c.1	NIP51	10	4	1
LGA.d.1	LRL14	9	0	0
LGA.d.2	LRL14	36	24	5
LGA.d.3	LRL27	2	2	2
LGA.e.1	K1468	1	1	1
LGA.e.2	SCP17	6	2	1
LGB.a.1	LRL14	29	11	5
LGB.c.2	DTX27	0	0	0
LGB.d.2	PSMD4	2	0	0
LGB.e.4	FB311	0	0	0
LGB.e.6	EDR1	3	0	0
LGC.a.1	ML328	10	4	1
LGD.a.1	EPN1	0	0	0
LGD.a.3	PER1	1	0	0
LGD.b.1	PME61	6	3	2
LGD.b.3	CRK3	9	0	0
LGE.a.3	GLIP2	0	0	0
LGF.a.1	PWP2	1	0	0
LGF.a.2	PWP2	2	1	0
LGF.b.1	PPP7L	0	0	0
LGG.b.1	NIC1	0	0	0
LGG.c.2	CXE5	9	1	1
LGG.d.1	ERF92	0	0	0
LGJ.a.1	PRU1	2	2	2
LGJ.a.2	BEV1D	4	2	1
LGK.a.1	MIEL1	1	1	1
LGL.a.2	PPP7L	1	1	1
LGL.c.2	DRL42	0	0	0
LGL.c.5	DRL42	3	0	0
LGL.d.1	POLX	0	0	0

Table 3.16 List of predicted genes in regions where most blight-associated polymorphisms were found showing predicted function and evidence for association with blight resistance based on statistical association and publicly available cDNA data.

Locus	Gene[a]	Exon[b]	NSyn[c]	Inferred function[d]	Transcript[e]	Diff[f]	Clust[g]
LGA.a	g1418	3	2	Aberrant root formation protein 4	CC 2, AC 1	na	1
LGA.b	g2361	0	0	Lysophospholipase	CC 2, AC 3	na	0
LGA.c	g3528	1	0	NIP5-like aquaporin	CC 1, AC 2	CC	0
LGA.d	g4191	0	0	LRK10-like rust resistance	AC 1	na	3
LGA.d	g4193	0	0	LRK10-like rust resistance	CC 4, AC 3	AC	3
LGA.d	g4196	4	3	LRK10-like rust resistance	CC 1	na	3
LGA.e	g8459	19	7	LISH/HEAT-domain protein	CC 7, AC 1	CC	0
LGA.e	g8465	5	1	Serine carboxypeptidase	CC 1	CC	0
LGB.a	g2160	16	9	TAO1-like TMV resistance protein	na	na	0
LGB.b	g2214	0	0	Protein PIN-LIKES 5	CC 2, AC 1	CC	0
LGB.b	g2245	1	1	Cytochrome P450 90B1	CC 1, AC 1	na	0
LGB.c	g3006	2	1	DETOXIFICATION 27 MATE-like	na	na	2
LGB.d	g5043	1	1	F-box protein	CC 1	na	6
LGB.d	g5048	2	1	Protein kinase EDR1	AC 1	AC	0
LGC.a	g3384	0	0	MLP-like protein 328	CC1	na	2
LGC.a	g3419	0	0	LRK10-like rust resistance	CC1	na	0
LGD.a	g1162	0	0	EARLY FLOWERING 3 -like	CC 4, AC 1	AC	0
LGD.a	g1179	2	1	Cationic peroxidase	CC 1, AC 1	na	0
LGD.b	g2262	0	0	Pectinesterase inhibitor	CC 1	na	0
LGD.b	g2282	3	1	Cysteine-rich RLK	CC 1, AC 1	na	1
LGE.a	g7940	0	0	GDSL esterase-lipase 2-like	CC1, AC 1	na	0
LGF.a	g1803	1	0	Periodic tryptophan protein	AC 1	na	1
LGF.a	g1804	1	0	Periodic tryptophan protein	CC 1	na	1
LGG.a	g2311	1	0	Senescence/dehydration-associated	AC 2, CC 2	na	0
LGG.b	g3657	0	0	Nicotinamidase 1	AC 1, CC 2	CC	0
LGG.c	g4295	0	0	Probable carboxylesterase 5	AC 3, CC 1	CC	5
LGG.c	g4298	2	0	2-hydroxyisoflavanone dehydratase	AC 1, CC 1	na	2
LGJ.a	g238	0	0	Pathogenesis-related protein	AC 1, CC 2	na	2
LGJ.a	g240	4	3	Cytosolic carboxypeptidase	na	na	0
LGJ.b	g1363	0	0	FLX-like protein	AC 1, CC 2	na	0
LGK.a	g2007	0	0	MIEL1 ubiquitin-protein ligase	CC 1	CC	0

Table 3.16 continued

Locus	Gene[a]	Exon[b]	NSyn[c]	Inferred function[d]	Transcript[e]	Diff[f]	Clust[g]
LGL.a	g4222	9	5	MAIN-like protein phosphatase	na	na	2
LGL.b	g6971	0	0	Probable disease resistance protein	AC 3, CC 2	AC	10
LGL.b	g6992	2	0	Probable disease resistance protein	CC 2	na	10
LGL.c	g8955	0	0	Retrovirus-related POL polyprotein	AC 1	AC	0

[a] Number assigned to predicted gene by AUGUSTUS gene prediction software; [b] Number of polymorphisms in predicted exons with Plink association p-value < 0.01; [c] SNPs predicted to cause an amino acid change with Plink association p-value <0.01; [d] Function inferred from alignment to SwissProt/UniProt database; [e] Number of cDNA contigs from Barakat et al. (2012) matching predicted protein (>75% ID) in American (AC) and Chinese (CC) chestnut; [f] Differential expression in cankers vs. healthy stem tissue in American (AC) and Chinese (CC) chestnut (Barakat et al. 2012); [g] Size of gene cluster, i.e. number of genes with same or similar predicted function adjacent to the named gene

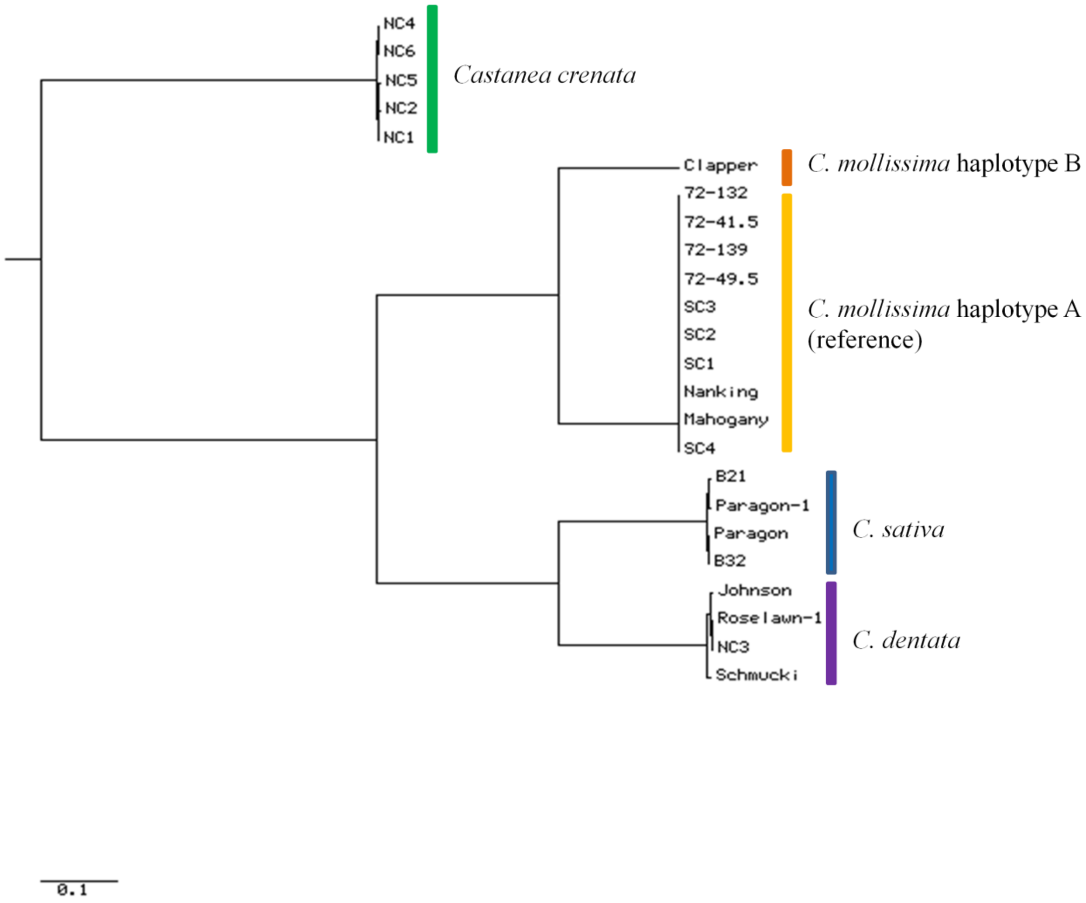


Figure 3.1 Maximum-likelihood tree constructed using SNP polymorphisms from assembled chloroplast genomes of 24 chestnut samples, showing two distinct chloroplast haplotypes of *Castanea mollissima*, one *Castanea dentata* haplotype, a *Castanea sativa* haplotype from “Paragon” in its offspring, and a *Castanea crenata* haplotype in Korean-derived *C. mollissima* material.

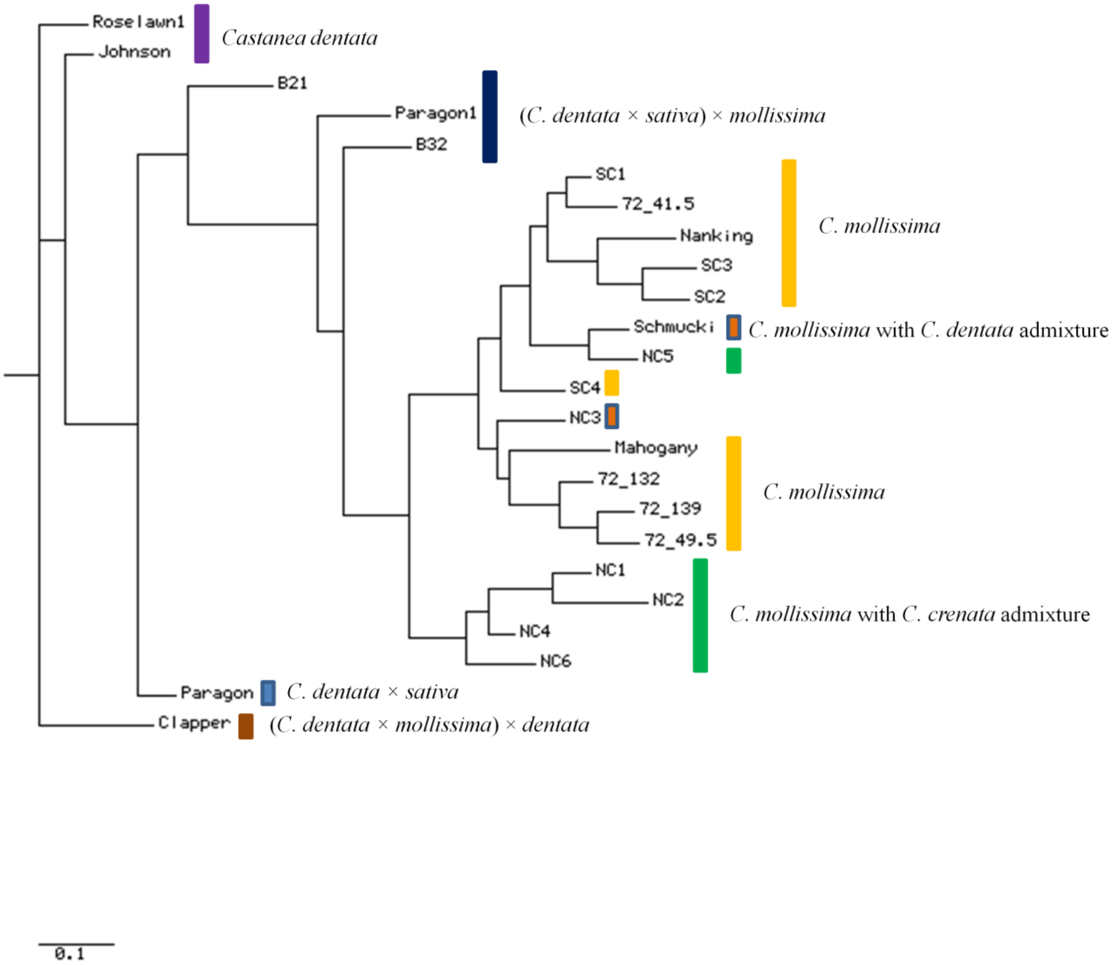


Figure 3.2 Maximum-likelihood tree (SNPhylo) constructed using SNP polymorphisms in predicted exons of linkage group A (LGA) pseudochromosome assemblies.

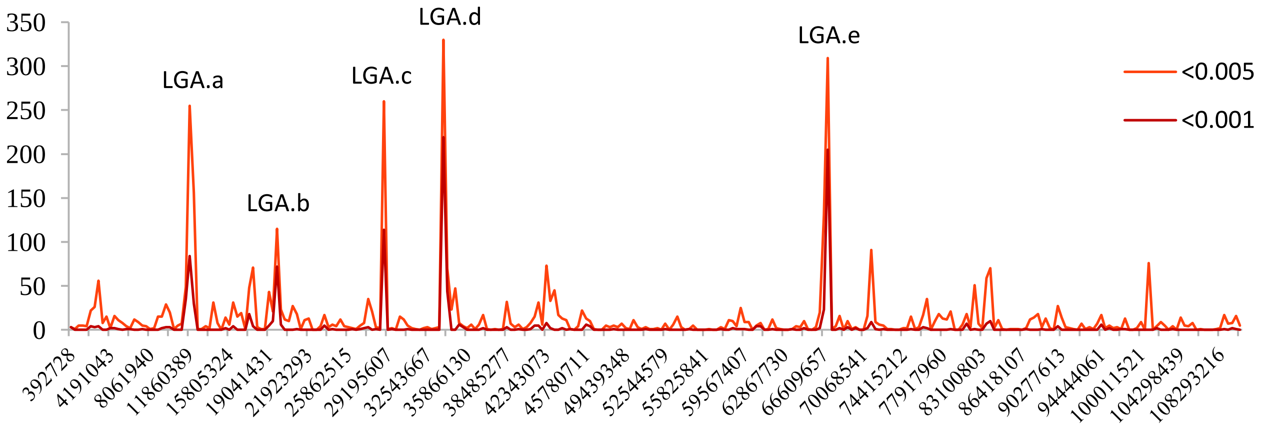


Figure 3.3 Association results for LGA, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

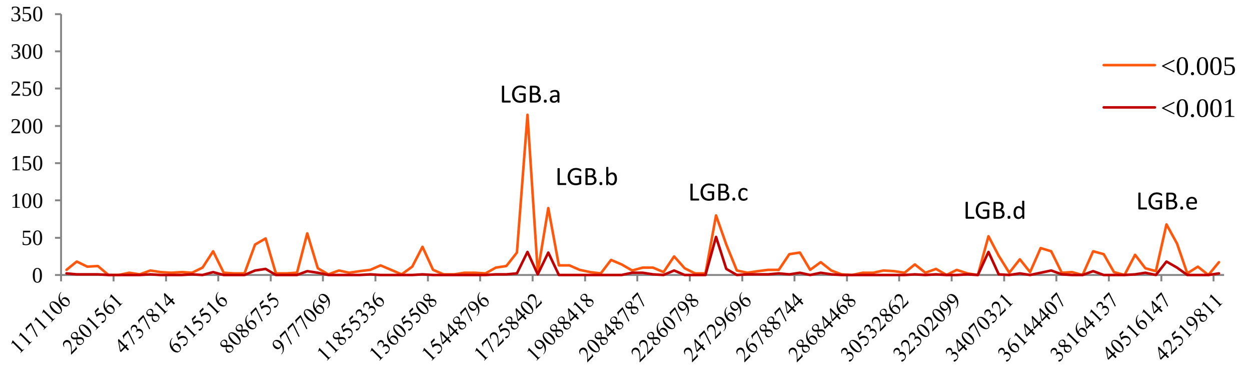


Figure 3.4 Association results for LGB, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

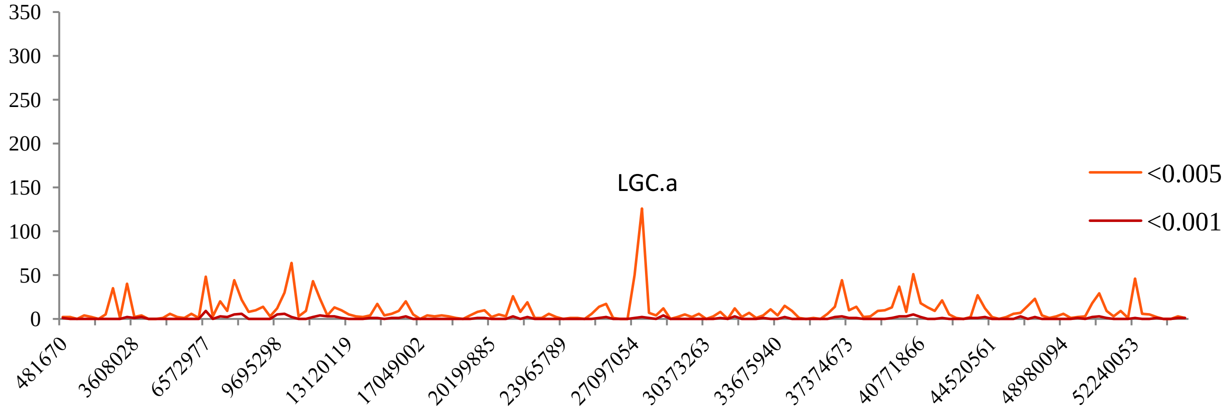


Figure 3.5 Association results for LGC, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

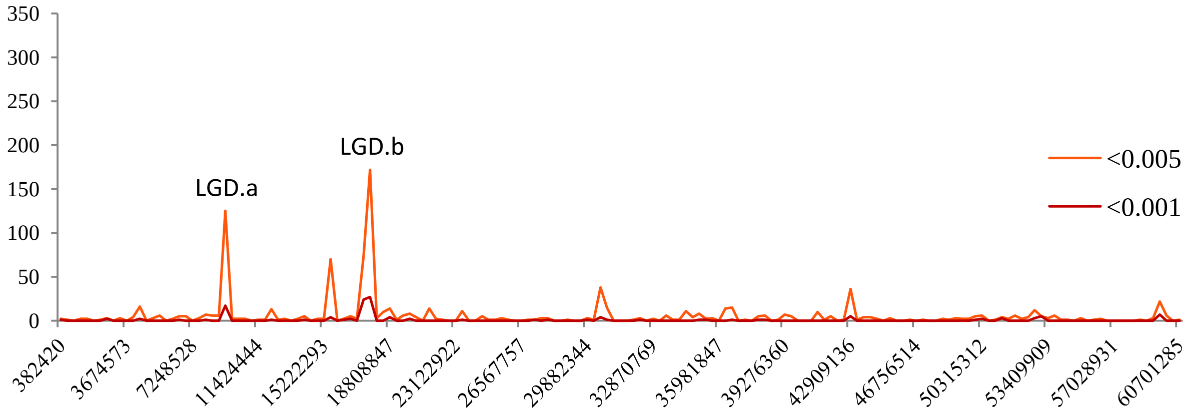


Figure 3.6 Association results for LGD, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

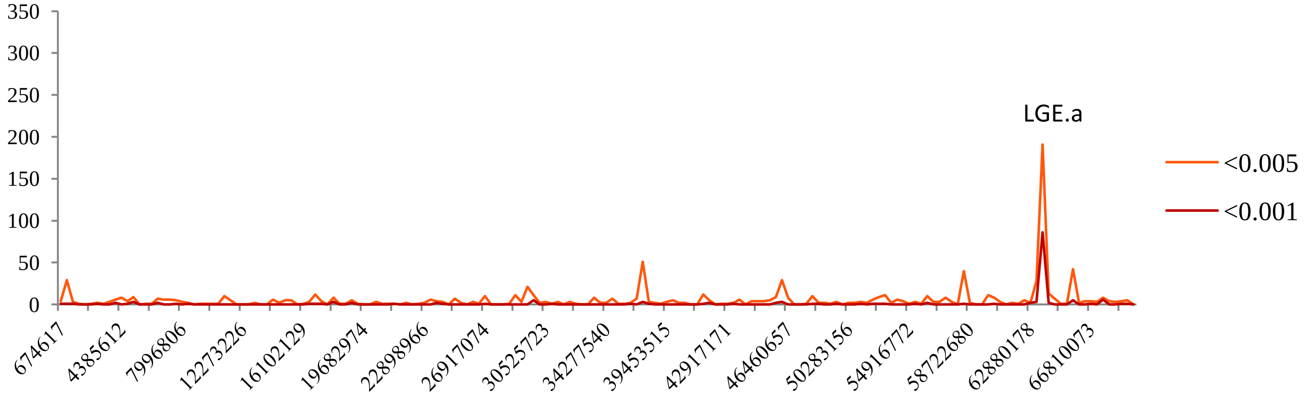


Figure 3.7 Association results for LGE, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

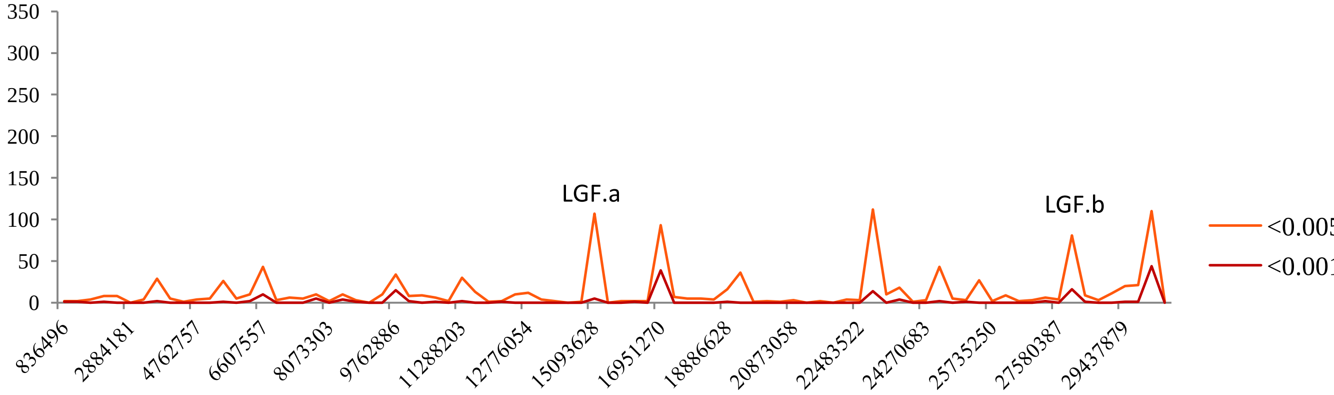


Figure 3.8 Association results for LGF, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

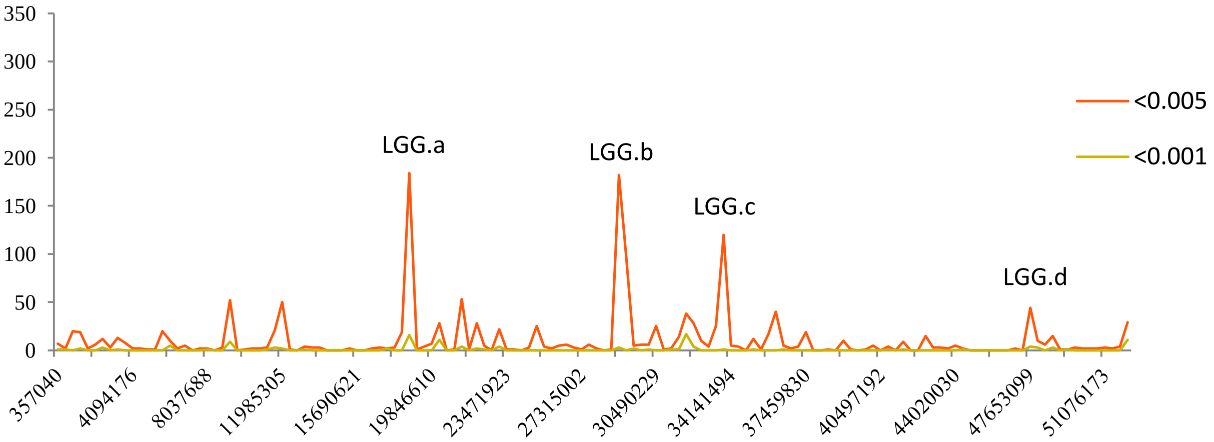


Figure 3.9 Association results for LGG, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

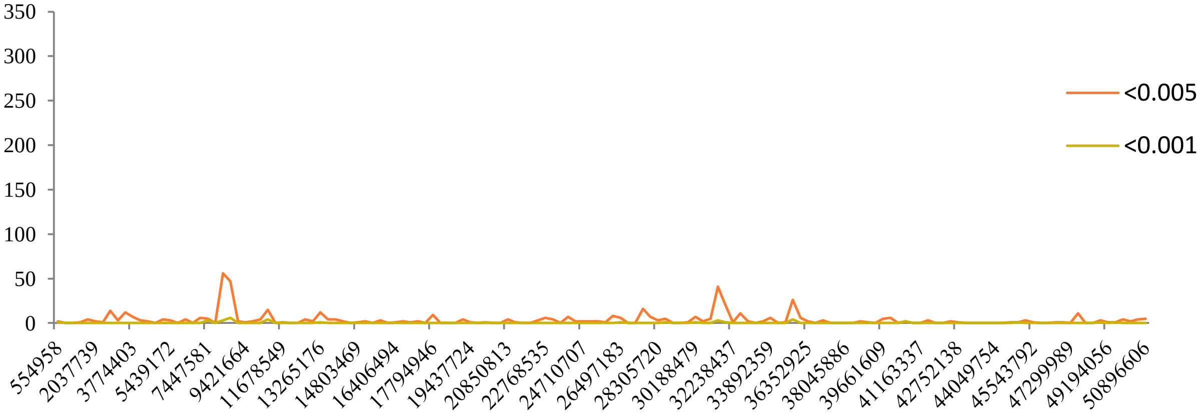


Figure 3.10 Association results for LGH, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

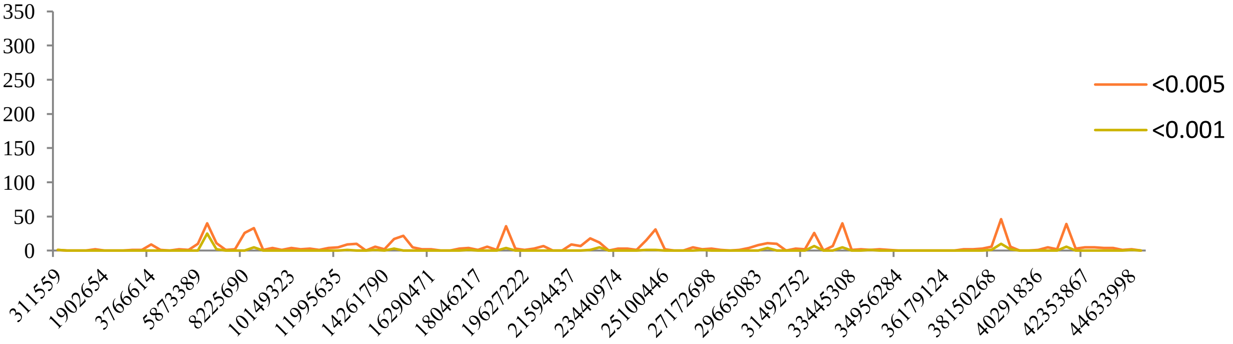


Figure 3.11 Association results for LGI, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

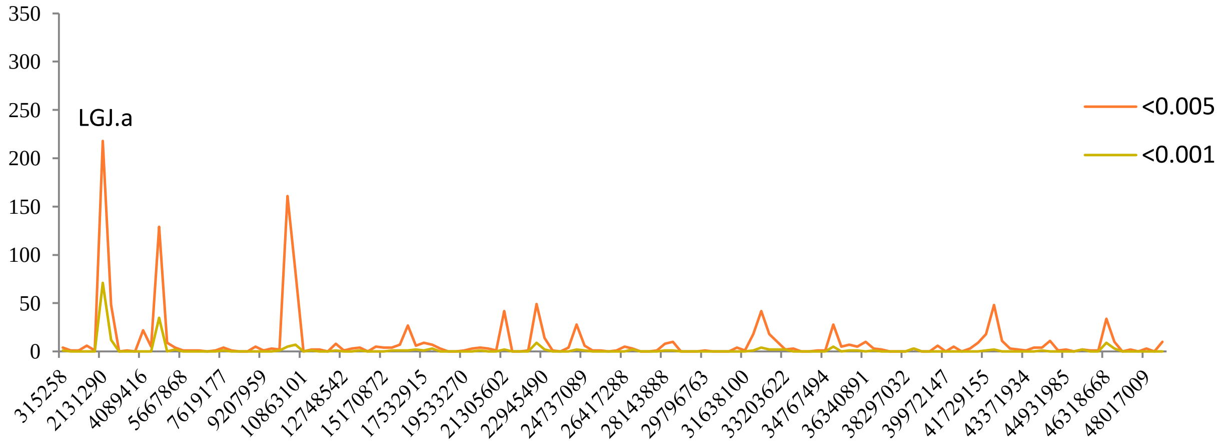


Figure 3.12 Association results for LGJ, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

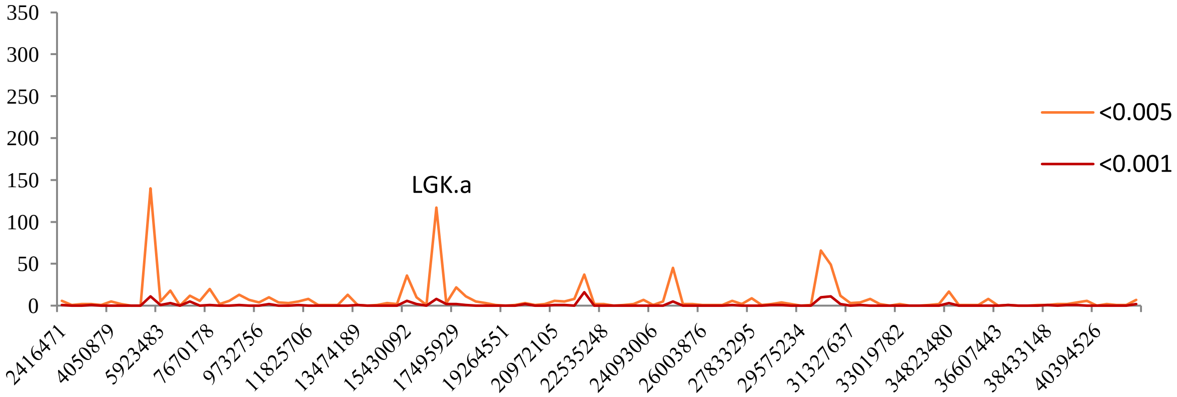


Figure 3.13 Association results for LGK, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis.

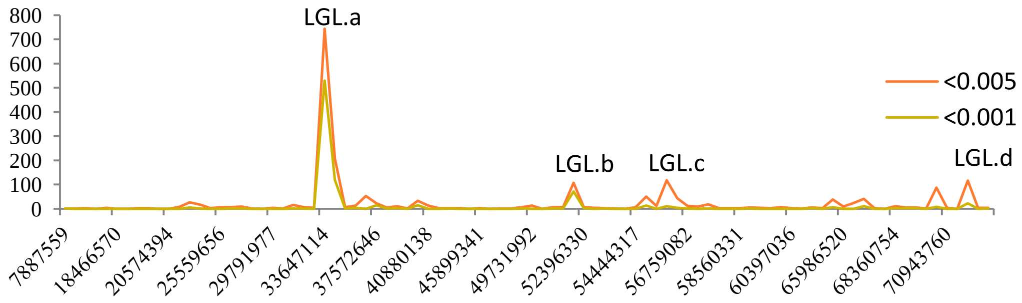


Figure 3.14 Association results for LGL, with bp on the horizontal axis and number of associated SNPs per 5000 on the vertical axis. Note that the vertical axis reaches larger values for this linkage group.

CHAPTER 4. IDENTIFICATION OF LOCI IN THE GENOMES OF BACKCROSSED HYBRID CHESTNUT (*CASTANEA*) THAT INFLUENCE CACHING DECISIONS OF FOX SQUIRRELS (*SCIURUS NIGER* L.)

Abstract

Dispersal of seeds by scatter-hoarding rodents is a strategy common to many tropical and deciduous tree species, notably in trees of the order Fagales (beech, oak, chestnut, walnut, and hickory) that define many North American forest ecosystems. I used tagged seeds to measure average dispersal distance for seeds of backcrossed hybrid chestnut (*Castanea*). Since the parent species (*Castanea dentata* and *C. mollissima*) have distinct seed phenotypes and tend to be dispersed different distances, the variation in seed traits and dispersal distance observed is likely to be caused by genetic variation among parent species that is unevenly distributed among backcrossed hybrids. To identify candidate genomic regions, I scanned genome sequences for pools of mother trees with variable dispersal measurements (high caching rate/long distance; low caching rate/short distance; no caching). Candidate regions for seed dispersal were identified as loci with more *Castanea mollissima* alleles in the high caching rate/ long distance pool than expected by chance and observed in the other two pools. These regions contained predicted lipid metabolism, dormancy regulation, seed development, and carbohydrate metabolism genes that have plausible roles in influencing seed caching behavior by tree squirrels.

4. 1 Introduction

Successful dispersal of seeds is an important facet of tree biology; effective dispersal allows parent trees to propagate within an immediate area, may decrease density-dependent mortality near parent trees, and enables a tree’s offspring to colonize suitable habitat farther away (Janzen 1971, Howe and Smallwood 1982). Long-distance dispersal allows trees to expand their range, which reduces the risk of extinction (Vander Wall 2001, Schupp et al. 2010, Larson-Johnson 2015). Among angiosperms, diverse strategies for seed dispersal have evolved (Friis et al. 2011). In the angiosperm-dominated forests of eastern North America, dominant canopy species produce seeds with wings of varying sizes (*Acer, Fraxinus*, *Betula*, *Liriodendron*) and tufts (*Populus*) for wind dispersal, seeds surrounded by a soft pulp (*Prunus, Celtis*) that provide attractive food for animals and are dispersed after passing through the digestive tract, and large nutritious seeds surrounded by a tough shell (*Quercus, Fagus, Castanea*, *Carya*, *Juglans*) that are primarily dispersed by animals that cache seeds in the ground (Leopold et al. 1998). The latter dispersal strategy, in North America, is mostly restricted to members of the order Fagales, which includes a large number of nut-bearing, animal dispersed taxa (*Quercus, Juglans*, *Corylus*) as well as wind-dispersed taxa (*Betula, Alnus*) (Vander Wall 2001). Oaks (*Quercus*), hickories (*Carya*) and formerly chestnuts (*Castanea*), which produce large animal-dispersed seeds, dominate diverse forest ecosystems over large areas of the eastern and central United States (Leopold et al. 1998). In the order Fagales, dispersal by seed caching is associated with increased range size and greater species diversity in animal-dispersed genera (Larson-Johnson 2015). Across many tree taxa, dispersal by seed-caching animals is actually associated with greater dispersal distance than wind dispersal (Thomson et al. 2011).

The first nuts are believed to have evolved from a winged nutlet similar to those seen today in extant Fagales taxa such as *Pterocarya* (Vander Wall 2001). The primary advantage of a large seed is that it provides the seedling with a large store of water, carbon, and nutrients with which to rapidly form effective roots and shoots (Gomez 2002). Since small seeds come with limited nutrient reserves, they require a constant supply of moisture to establish a root system and form a shoot to begin photosynthesis, which gives an advantage to larger seeds in dry environments (Larios et al. 2014). Dry climatic conditions are thought to have been the selective pressure that originally drove the evolution of angiosperm nuts, which first appeared in the Paleocene (66-55 MYA) (Vander Wall 2000). Large seeds can also confer an advantage in shaded environments, where the development of large leaves quickly from seed reserves can make seedlings more competitive and enable survival in low-light environments (Baker 1972, Lohbeck et al. 2015).

Distinct disadvantages are also associated with large seeds. The large nutrient investment in each individual seed means that fewer seeds overall can be produced

(Venable 1992). Their large mass means that passive dispersal and wind dispersal are less likely to remove seeds beyond the canopy of the parent tree, although tree height may compensate for this potential disadvantage (Thomson et al. 2011). Finally, the large size and nutrient content of the seeds themselves make them an attractive food source for seed-eating insects and other animals, potentially leading to the destruction of large numbers of valuable seeds (Paulsen et al. 2014). Interaction with animals as seed predators is thought to have influenced the evolution of fruits and seeds in angiosperms at an early stage (Friis et al. 2011) and most nut-bearing trees have coevolved a "conditional mutualism" (Thiemer 2005) with rodents and birds that consume large numbers of nuts, but cache enough in the soil—sometimes at considerable distances from the mother tree—to allow a few to escape predation and germinate.

The larger-seeded members of Fagaceae and Juglandaceae are primarily dispersed by rodents (Vander Wall 2000). Coevolution with rodent conditional mutualists has led nut-producing trees to adopt a wide range of seed-packaging strategies. Some have hard and thick shells that can only be penetrated by rodents with powerful jaws and specialized teeth, like squirrels and some mice (Tamura et al. 2008). Some have extremely large seeds (e.g. black walnut; *Juglans nigra*) that can be readily handled only by large tree squirrels (*Sciurus* spp) which scatter-hoard seeds and therefore provide more effective dispersal than smaller rodents (Stapanian and Smith1978). Others produce small seeds (e.g. shingle oak, *Quercus imbricaria*) that can potentially dispersed by scatter-hoarding birds (e.g. blue jay, *Cyanocitta cristata*) expanding the number of potential dispersers and the dispersal range (Johnson and Webb 1989, Richardson et al. 2013). Smaller seeds, however, might be more subject to consumption if they are attractive to smaller rodents, such as the eastern chipmunk (*Tamias striatus*) (Blythe et al. 2015), that practice larder-hoarding. In larder-hoarding, many seeds are stashed in cavities and burrows that preclude germination or successful establishment of seedlings (Clarke and Kramer 1994). Some seed predators that do not provide effective dispersal, however, may prefer to consume larger seeds, such as the European wild boar (Gomez 2002).

Scatter-hoarding in the eastern North American deciduous forest is primarily practiced by two species of tree squirrels, the eastern gray squirrel (*Sciurus carolinensis*) and the fox squirrel (*Sciurus niger*), as well as the blue jay (*Cyanocitta cristata*, Steele et al. 2001, Smith and Stapanian 2002, Moore et al. 2007, Richardson et al. 2013, Blythe et al. 2015) . Seeds of chestnut and other Fagales trees are also predated by smaller squirrels (*Tamias striatus, Tamiasciurus hudsonicus*) and mice (*Peromyscus* spp), which are generally larder-hoarders, and larger animals like deer and turkeys which do not hoard seeds at all (Ivan and Swihart 2000, Steele et al. 2001, Goheen and Swihart 2003). Both larder-hoarding and scatter-hoarding require a large investment of time in locating, moving and caching seeds, but a successful scatter-hoarder must also be able to relocate scatter-hoarded seeds and expose itself to predators in the process, which a larder-hoarder does not need to do (Steele et al. 2014). A larder-hoarder, however, risks losing its entire food supply if the larder is pilfered by another animal; the advantage of scatter-hoarding is that the squirrel is free to engage in activities other than guarding, including foraging for more food (Stapanian and Smith 1978, Vander Wall 2001, Moore et al. 2007, Brodin 2010) instead of vigilantly defending its larder. While a large number of individual scattered caches may be pilfered, it is unlikely that all of them will be (Vander Wall and Jenkins 2003). In general, the consumption rate of cached seeds is quite high: rodents appear to use a combination of spatial memory and olfaction to detect seeds they have cached, and most of the seeds not consumed by the original individual hoarder are dug and eaten either by conspecific cache-pilferers or by individuals of other species (Brodin 2010). Thompson and Thompson (1980) estimated that, of 500 artificially buried horsechestnuts (*Aesculus*), 85% were removed by rodents, and 12% ultimately germinated. Calhane (1942) found that fox squirrels (*Sciurus niger rufiventer*) in an urban cemetery removed 99% of naturally cached hickory (*Carya)* and white oak (*Quercus alba*) seeds and 86% of artificially cached seeds. In a sample of 240 Japanese walnut (*Juglans ailantifolia*) seeds cached by Japanese squirrels (*Sciurus lis*) observed over 30 days by Tamura et al (1999), 80% were consumed either by the squirrels or by wood mice (*Apodemus speciosus*). Because so few seeds survive the caching process, there must be some germination advantage for cached vs. non-cached seeds in order for trees and squirrels to mutually benefit from scatter-hoarding (Zwolak and Crone 2012). Caching in the soil hides the seed from other potential predators at the surface and can prevent desiccation (Vander Wall 2005, Schupp 2010, Zwolak and Crone 2011). In thin-shelled nuts like oak and chestnut, death through desiccation is a danger for exposed seeds (Connor et al. 2006), so these species are likely to derive a net benefit from their relationship with scatter-hoarding rodents.

If the American chestnut (*Castanea dentata*) (abbreviated *Cd* elsewhere in this document) is restored to the eastern forest landscape, it will likely be planted in reclaimed minelands, old fields, and other open sites (Jacobs 2007) and will need to subsequently re-integrate itself into the forest with the help of its co-evolved seed dispersers, the scatter-hoarding tree squirrels (Jacobs et al. 2012). If the backcrossed hybrid trees produced by TACF are ultimately used for restoration, they will include some genetic material from Chinese chestnut (*Castanea mollissima*) (abbreviated *Cm* elsewhere in this document) that, in addition to conferring enhanced blight resistance, may include genes that influence seed traits (Worthen et al. 2010). There is evidence that seed of BC3 chestnuts, which only contain a small fraction of the *Cm* genome, show a dispersal pattern that isdistinct from *Cd*(Blythe et al. 2015). Since the seeds of hybrid-BC3 trees were dispersed farther on average than *Cd* in Blythe et al. (2015), hybrid background may not negatively affect the fitness of restored chestnut, but nevertheless represents a potential unintended effect of hybrid breeding on the ecological relationships of restored chestnut in the eastern North American hardwood forest. One impetus for the restoration of chestnut to the eastern North American deciduous forest is its unique ecological value, as chestnut produces large crops of thin-shelled, nutritious seeds on a nearly-annual basis (Dalgleish and Swihart 2012). The ecological function of hybrids should be considered in restoration plans. Furthermore, differences in seed traits might influence the viability of hybrids at the northern limits of the American chestnut’s native range, where Chinese chestnut is poorly adapted to long, cold winters (Saielli et al. 2012).

The difference in dispersal distance and caching vs. consumption rates between *Cd* and *Cd* x *Cm* nuts is probably due to a number of differences in the nuts of these tree species. Tree squirrels, including *Callosciurus erythraeus* and *Sciurotamias davidianus*, are important dispersal agents of *Cm* in its native range (Xiao et al. 2013), so the selective forces exerted by conditional seed dispersers on *Cm* seeds have likely been similar to those acting on *Cd*. *Cm* seeds are larger, on average, than American chestnuts, and squirrels tend to be more likely to cache a large seed than a small one, and tend to carry the seed a longer distance before caching (Jansen et al. 2002, Xiao et al. 2005). The variation in seed size among individual trees appears to have a strong genetic basis, and geographic variation can be observed in some species. In Japanese walnut (*Juglans ailantifolia*), gradients in nut size correlated with the dominant dispersal agent (squirrels vs. mice) in different geographic areas (Tamura and Hayashi 2008). Tendencies of North American oaks (*Quercus*) to produce smaller nuts in the northern parts of their ranges have led to the hypothesis that long-distance dispersal by birds caused over-representation of small-fruited trees as species advanced north from glacial refugia (Johnson and Webb 1989). Nut size is an important factor in a squirrel's caching decision, but it does not appear to be the sole consideration. Most nuts germinate in the spring after a dormant period over winter, but some trees, like white oak (*Quercus alba*) and many of its relatives, begin germination in the fall (Fox 1982, Xiao et al. 2009a,b). Once a seed begins to germinate, drastic physiological changes occur that reduce the seed's value as a food source. Starches are converted into sugars, sugars are hydrolyzed in respiration as new root and shoot tissues develop, and fat and protein reserves are converted to the raw materials of the growing root tip (Baskin and Baskin 1998). Squirrels are aware of these changes, and take measures to maximize the nutritional utility of non-dormant seeds (Fox 1982, Steele 2001, Smallwood et al. 2001). They are more likely to eat, rather than cache, non-dormant seeds, and have been observed removing the embryonic axis of a non-dormant seed prior to caching, so that it cannot germinate (Smallwood et al. 2001). Conversely, a dormant seed is more likely to be cached and carried a longer distance. Chestnuts go through true dormancy with a chilling period (Baskin and Baskin 1998), so any difference in dormancy-breaking behavior among *Cd*, *Cm*, and hybrids would be marginal, and presumably have a small influence on seed disperser behavior. However, physiological differences in the seeds of different chestnut species could be interpreted by squirrels as signals of dormancy or impending germination, in particular, different levels of tannins, differences in the wax layers of the pericarp, volatile organic compounds released from the nut, and differences desiccation rate that could be influenced by pericarp thickness and waxy coatings (Steele et al. 2001, Sundaram 2016). Desiccation causes changes in the flavor and quality of chestnut cotyledons that are perceptible to humans (Rutter et al. 1991, Ertan et al. 2015). If squirrels interpret these signals as pertaining to germination and the breaking of dormancy, they could influence caching behavior (Steele et al. 2001).

It is likely that many of the traits that influence seed dispersal are heritable, given that natural selection has acted on them during the coevolution of trees and seed dispersers, and that similar traits (seed size, seed nutritional makeup, and volatile organic compounds) have been selected successfully by humans in plant breeding programs. Given the large size of *Cm* seeds relative to *Cd*, it is likely that *Cm* possess alleles at seed trait-related genes that are highly divergent from alleles in *Cd.* Other genes that influence seed dispersal might be similarly divergent between species. Individual BC3 trees vary widely in seed size, pericarp color, and other seed traits; these differences may be due to differences in the *Cm* alleles individual BC3 trees inherit rather than differences in their *Cd* parentage. The phenotypes of interspecific hybrids are not necessarily intermediate between parent species (Woeste et al. 1998). For those BC3s whose seed traits, especially size, fall outside the normal range of variability for *Cd*, it is likely that inheritance of seed trait alleles from *Cm* is causative of differences in seed morphology and dispersal. Since the trees in BC3 populations have mixed parentage and do not constitute a true QTL mapping population, conventional QTL mapping would not be the best way to uncover loci underlying differences in seed traits (and seed dispersal) among the sizable population of BC3 chestnuts in blight-resistance screening orchards. But, since the differences are due to inheritance from a different species (*Cm*) in the *Cd* genomic background of BC3s, it should be possible to isolate loci that influence seed traits and dispersal by associating presence or absence of *Cm* alleles with dispersal and other seed traits.

The goal of our experiment was to determine if there are identifiable loci in the genomes of hybrid chestnut where a *Cd/Cm*, versus a *Cd/Cd,* genotype is associated with differences in seed dispersal distance and likelihood of caching versus consumption of seeds. Our research questions were:

1) Do differences in seed size or other heritable characteristics influence differences in the way dispersers handle, consume, and/or cache backcrossed hybrid chestnut seeds?

2) What is the genetic basis of seed traits that lead to differences in seed disperser (squirrel) behavior during interactions with different individual hybrid chestnuts?

Based on previous work (Blythe et al. 2015, Sundaram et al. 2016), we predict that squirrels will tend to disperse seeds that are larger (more similar to Chinese chestnut) farther than seeds that are smaller (more similar to American chestnut) and they will cache them more frequently relative to the number of seeds consumed without caching, and that the mother trees of seeds that are dispersed farther and cached more frequently have a *Cm/Cd* genotype at loci in the genome that contain genes predicted to have a plausible association with seed traits that are related to dispersal. To test these hypotheses, we collected seeds from BC3 chestnuts growing at Purdue University, obtained phenotypes for seed weight, dispersal frequency, and dispersal distance over 3 years.

4.2 Materials and Methods

4.2.1 Seed collection

Seeds were collected in late September and early October in a planting of BC3([(*Castanea mollissima* × *dentata)* × *dentata*] × *dentata*) × *dentata*) chestnuts at Purdue University's Lugar Farm in Tippecanoe County, IN. The purpose of the planting is selection for blight resistance with the ultimate goal of restoring American chestnut to southern Indiana forests. Most of the seed parents were from a section of the planting that was 11 years old in 2014, with some also from a younger section that was 4 years old in 2014 at the start of the study. 'Clapper,' a BC1 tree, was the blight-resistance donor and only source of *Cm* genetic material in this backcross population. *Cm* nuts were obtained from a pair of trees planted as blight-resistant checks in the Lugar Farm orchards. *Cd* nuts were obtained from two adult trees growing at the Purdue Wildlife Area in Tippecanoe County, IN. BC3 seed parents were chosen based on seed size, with roughly equal numbers of large-seeded, small-seeded, and average-seeded trees chosen. Seed parents were tagged with durable individual plastic nursery labels, but due to annual variation in the size of seed crops, and declining health of some seed parents due to a severe chestnut blight infestation in the planting, different seed parents were chosen each year of the study. Seeds were collected by knocking burrs off the parent tree using a ~ 2 m wooden pole and manually removing seeds from the bur if necessary. Seeds were floated in water to determine viability; floating seeds were deemed non-viable and discarded.

4.2.2 Seed measurements

Seeds were stored in a cooler (40 degrees F) following de-burring and floating and stratified in peat moss to maintain viability during cold storage. In October, at least ten seeds from each seed parent were weighed on a digital scale to determine average seed mass. Length (from seed base to tip) and width (across the broadest part of the seed) were determined using a digital calipers. In 2015, desiccation was also measured by weighing seeds immediately after collection and again 80 days following collection.

4.2.3 Seed tagging

Tagging was carried out immediately before dispersal trials to avoid spoilage of seeds. A method similar to that employed by Xiao (2006) and Hirsch et al. (2012) was used. A hole was made in the proximal (wider) end of each seed using either a botanical dissecting needle or a small (~2 mm) drill bit. A piece of 24-gauge green floral wire approximately 12 cm long was looped through the hole and twisted to secure it. A piece of brightly colored waterproof tape was attached to the end of the wire and labeled with a number designating the seed parent.

4.2.4 Dispersal trials

Dispersal trials were carried out in November and December of each year at several feeding stations placed in and around the Lugar Farm chestnut plantings in Tippecanoe County, IN and in 2016 at one additional location in Woodford County, IL adjacent to the campus of Eureka College. At Lugar Farm, several hundred BC3 chestnuts were present with black walnut (*Juglans nigra*) abundant in adjacent woodlots and fencerows along with several oak species (primarily *Quercus palustris, Q. velutina,* and *Q. imbricaria*). In Eureka, fencerows and woodlots were not present at the testing site but large black walnuts, oaks (*Quercus muehlenbergii, Q. imbricaria, Q. velutina*) hickory (*Carya cordiformis*) and buckeye (*Aesculus flava*) and several *Cm* were all nut producers in the local suburban forest canopy. Fox squirrels (*Sciurus niger*) were the only scatter-hoarding squirrel species observed at each feeding station, although red squirrels (*Tamiasciurus hudsonicus*) were observed in pine (*Pinus strobus*) plantings at Lugar Farm. White-tailed deeer (*Odocoileus virginianus*) hoofprints and droppings were frequently observed near two feeding stations at Lugar Farm. Wild turkeys (*Meleagris gallopavo*) and their feathers were observed at one of the Lugar Farm feeding stations. Feeding stations were pre-baited with black walnuts, peanuts, corn, sunflower seeds and a mixture of peanut butter, molasses, and oatmeal to acclimate local squirrels to the feeding locations in August prior to dispersal trials. During dispersal trials, 10 (2016) or 25 (2014-15) seeds from 5-6 (2016) or 3-4 (2014-15) parent trees were randomly distributed near a post at the center of each feeding site. Seeds were left out for 4-5 days, and seed fates (cached, consumed, or left at feeding station) were recorded and dispersal distances measured with a forestry measuring tape attached to the post at the center of the feeding site. Intensive searches for seeds were conducted up to 20 m from the feeding site, although some seeds were found outside this distance. Trials started in late October or early November and continued through December until the soil surface froze. Relationships between seed dimensions and dispersal parameters were statistically investigated using the lm and glm packages of R software version 3.2.3 (R Core Team 2015).

4.2.5 DNA Isolation

DNA was isolated from BC3 seed parent trees following dispersal trials. Twigs were collected for DNA extraction in early spring 2016 and 2017. Terminal sections (about 7 cm) of 1st-year twigs were ground to a fine powder in liquid nitrogen using a mortar and pestle. The ground tissue was placed in 5 mL of heated (50 C) CTAB extraction buffer in a 15 mL conical tube and incubated 4-8 hours at 50 C. Following incubation, 1 mL of 20 mg/mL proteinase K solution was added and samples were incubated for an additional 15 minutes. 5 mL of 25:24:1 phenol:chloroform solution was added and samples were purified using a standard phenol:chloroform extraction (Doyle and Doyle 1987) followed by precipitation of DNA using 0.2 M sodium chloride and isopropanol. After pelletting and resuspending samples in TE buffer, contaminants were removed using Zymo Research OneStep PCR Inhibitor Removal kits (Zymo Research). Following purification, samples were quantified using a Nanodrop 8000 (ThermoFisher Scientific), and 2% agarose gel, pooled, then submitted to the Purdue Genomics Core Facility for sequencing.

4.2.6 DNA pooling and sequencing

Pools of samples were made for different phenotypic classes based on 1) mean dispersal distance for cached seeds and 2) frequency of caching. The strong-dispersal pool (Pool A; 8 samples) contained DNA from parents that produced seeds with a long dispersal distance and high frequency of caching (i.e., successful dispersal), including one Chinese chestnut. A moderate-dispersal pool (Pool B; 7 samples) contained parents that produced seeds with a shorter dispersal distance and low frequency of caching. The weak-dispersal pool (Pool C; 10 samples) contained seed parents that produced seeds with a frequency of caching and dispersal distance near 0, including one American chestnut (Table 1). 20 uL of DNA at concentration 200 ng/ uL from each sample were included in a pool and the combined sample was submitted for library construction and sequencing; samples were sequenced as separate libraries on one Illumina HiSeq 2500 (Illumina Inc., San Diego, CA) lane. Reads were paired-end, 100 bp in length. The individual genomes of "Clapper," several unrelated Chinese chestnuts, and two American chestnuts were sequenced separately with 2 samples per lane (Chapter 3).

4.2.7 Genome assembly and SNP calling

Short reads were assembled to a draft Chinese chestnut reference genome provided by Dr. John Carlson of Penn State using the Burrows-Wheeler aligner. Alignments were processed and polymorphisms called for each pool of samples using Picard Tools and the Genome Analysis ToolKit (GATK) best practices workflow, minus the quality score recalibration step. When calling SNPs using the HaplotypeCaller tool from GATK, ploidy was set to twice the number of individuals in the pool.

4.2.8 Analysis of SNP data

Since most SNP analysis tools cannot process polyploid variants, custom Perl scripts were used to analyze the data. Based on the observation that many genes have low rates of polymorphism and heterozygosity within species but high rates of allele-frequency divergence among chestnut species (Chapter 3), and the hypothesis that variation in dispersal among BC3s is due to variation in the amount of *Cm* ancestry in individual BC3 trees, the goal of these scripts was to discriminate between predicted genes that had two genotypes in roughly equal proportions in a pool (*Cm/Cd* sites) and genes that had a single genotype fixed or nearly fixed within a pool (*Cd/Cd* sites). Since all individuals genotyped were BC3s, there should be no *Cm/Cm* sites in their genomes. This was accomplished by filtering the genome SNP file for polymorphisms occurring within predicted genes (AUGUSTUS gene prediction; Stanke et al. 2006) with strong (e-value <0.001) alignments to the Uniprot/SwissProt protein database. For each SNP (minimum depth = 8) within the predicted transcription start and stop sites of each gene, a value (hybridity estimator or HE) was assigned to approximate the proportion of individuals that were heterozygotes within a pool. If major allele frequency for a SNP within a pool was between 0.45 and 0.55, HE was assigned a value of 0.75 (most individuals heterozygous), if it was between 0.55 and 0.70, HE was 0.5, if it was between 0.70 and 0.85, HE was assigned the value 0.25, and if the major AF was >0.85 HE was assigned 0 (all or nearly all individuals homozygous) This estimate was averaged across all the SNPs in each predicted gene sequence; if more than 50% of SNPs in a gene sequence were missing genotypes, the gene was assigned a missing value. Finally, the hybrid estimate for genes was averaged in 10-gene bins for each pool and compared among pools. 10-gene bins that had a difference in average HE values >2 standard deviations greater than the average HE difference between the strong-dispersal pool and the moderate- and weak-dispersal pools were identified as candidate loci potentially contributing to differences in seed dispersal. To test the accuracy of the heterozygosity estimate, the estimated heterozygosity values averaged over the three pools was analyzed as a predictor variable for the heterozygosity values from the Clapper whole genome sequence using a simple linear regression and the first 268 genes from the linkage group A (LGA) pseudochromosome sequence. Genes within these regions were annotated using the UniProt entries for aligned proteins from the UniProt KB/ SwissProt database. Predicted molecular interactions were analyzed using the STRING protein database.

All the BC3 trees in our sample inherited 100% of their *Cm* alleles from ‘Clapper.’ ~50% of the genome of ‘Clapper’ (a BC1 tree) consists of loci with *Cd/Cm* genotypes. Of loci that are hybrid in ‘Clapper’, a given BC3 grandchild of ‘Clapper’ is expected to retain one of four loci as *Cm/Cd* with the rest switching to *Cd/Cd* due to recombination. Therefore, if a locus is chosen at random from among loci known to have a *Cd/Cm* genotype in ‘Clapper,’ a random sample of ‘Clapper’-derived BC3s is expected to have one *Cm* allele observed out of every eight. Since we genotyped BC3s in pools, opportunities for random sampling error were present; a given individual’s genotype might be over-represented at a locus, biasing allele frequency (and therefore, hybridity) estimates. Over-representation of an individual could be due to differences in DNA quality, inaccurate estimates of DNA concentration prior to pooling, or random inclusion of more DNA fragments from one individual during the high-throughput sequencing process. We developed a Perl script to estimate the likelihood that more *Cm* alleles than expected by chance alone were present at a given SNP locus in the pooled data.

First, a panel of eight *Cm* with no evidence of hybrid background, two *Cd*, and ‘Clapper’ whole-genome sequences (Chapter 3) were used to filter a whole-genome SNP genotype file for loci with one allele fixed in both *Cd*, one allele fixed in all eight *Cm*, and a *Cm/Cd* genotype in ‘Clapper’. The coordinates of these loci were recorded as informative SNPs. In the final step, these coordinates were used to determine which SNPs from the pooled data should be kept for analysis.

Next, the program made random draws from arrays of 100 values (0 for Cd, 1 for Cm) set to represent the expected species allele frequency at a given SNP locus for the strongstrong-dispersal, moderate-dispersal, and weak-dispersal pools. Since the strong-dispersal pool contained one *Cm*, the expected frequency of *Cm* alleles at a locus that was hybrid in ‘Clapper’ was 3/8 rather than 1/8; therefore, the array of potential alleles contained 38 “1” values and 62 “0” values. The moderate-dispersal pool only contained BC3s, so 1/8 was the expected fraction of Cm alleles. Since the weak-dispersal pool contained one *Cd*, the expected fraction of *Cm* alleles was slightly lower (1/11). To simulate the process of pooled DNA sequencing, random draws were made from this distribution up to a simulated read depth of 8 and the number of *Cm* alleles in the sample was tallied. This process was repeated 1,000,000 times for each pool to create null distributions for *Cm* allele frequencies in BC3 trees at ‘Clapper’ hybrid loci.

Finally, the SNP files for each pool were read, ignoring SNPs not matching the coordinates in the informative SNPs list from the first step. A p-value was assigned for each SNP in a pool based on the percent of simulated SNP genotypes that had a count of *Cm* alleles greater than or equal to the observed value. If this percentile-based p-value was lower than 0.05, the null hypothesis that *Cm* alleles were randomly distributed at the locus in a pool was rejected; low p-values were interpreted as evidence that more *Cm* alleles were present in the pool than expected by chance alone. For each predicted gene sequence in the genome, an average p-value was computed using all informative SNPs within the predicted gene sequence. Predicted genes where the null hypothesis was rejected in the strong-dispersal, but not in the moderate- or weak-dispersal pools were included as potential candidates for influencing seed dispersal.

4.2.9 Validation of predicted genes

To validate the predicted genes from the whole-genome analysis, cDNA data for a number of species in the order Fagales, the main nut-bearing taxon in the temperate zone, were aligned to predicted proteins from the *Castanea mollissima* genome. *Castanea mollissima* and *C. dentata* (Barakat et al. 2009, 2012; primarily stem tissue transcripts with some flowers and roots; downloaded from http://hardwoodgenomics.org/transcriptomes), *C. crenata* and *C. sativa* (Serrazina et al. 2015; roots only; downloaded from http://hardwoodgenomics.org/transcriptomes), *Quercus alba, Q. rubra, Fagus grandifolia, Alnus rubra, Alnus rhombifolia* and *Juglans nigra* (Hardwood Genomics Project; mixed tissues; downloaded from http://hardwoodgenomics.org/transcriptomes), *Quercus robur* and *Q. petraea* (Lesur et al. 2015, downloaded from https://arachne.pierroton.inra.fr/QuercusPortal/), *Juglans regia* (Martinez-Garcia et al. 2016; downloaded from http://dendrome.ucdavis.edu), *Corylus avellana* (Rowley et al. 2014; all tissues, downloaded from http://www.cavellanagenomeportal.com/), *Fagus sylvatica* and *F. crenata* (Ueno et al. 2009, Schlinck 2009, Lesur et al. 2015; downloaded from GenBank), *Betula platyphylla* (Mu et al. 2012) and *Nothofagus nervosa* (Torales et al. 2012; leaf library; downloaded from GenBank). cDNA contig consensus sequences were aligned to a database of predicted *Castanea* protein sequences using the Diamond sequence aligner (Buchfink et al. 2015). A predicted gene was counted as having transcript support if at least one cDNA contig had the predicted gene’s protein sequence as its best alignment.

4.3 Results

4.3.1 Seed phenotypes

Seeds were collected, and dispersal phenotypes obtained, for 13 BC3, American, and Chinese chestnut in 2014, 11 BC3, 1 American, and 1 Chinese chestnut in 2015, and 12 BC3 in 2016 (Table 1). The average mass (mean ± standard deviation) of BC3 seed over the three years was 3.51 ± 1.47 g, ranging between 1.12 and 7.78 g. The average for American chestnut was 3.05 ± 0.17; for Chinese chestnut the average was 7.82 ± 1.01. Average length from attachment point to tip of BC3 in 2014 and 2015 was 22.12 mm ± 2.43, with a range of 17.73-25.86 mm; length for American chestnut was 20.49 +/- 0.16, and for Chinese chestnut average length was 24.27±0.69. Width across the wider axis of the attachment-scar end of the nut was 20.63 ± 3.88 mm for BC3, ranging between 14.42 and 26.4 mm; for American chestnut the mean was 20.04 ± 0.70 mm and for Chinese chestnut 27.29 ± 0.46 mm. In seeds for which moisture loss was measured in 2015, *Cd* seeds lost more of their mass through drying (15.85%) than *Cm* (10.26%) under cold-room storage. The individual half-sib seed lots with the highest rate of caching (68, 55, and 25% of seeds recovered in caches rather than recovered eaten) lost moisture at rates similar to Chinese chestnut (8.99, 8.55, and 11.61% of moisture lost, respectively) while seeds that were less likely to be dispersed had highly variable (5.41-31.64%) loss of mass due to drying and an average rate of moisture loss (17.11%) closer to American than to Chinese chestnut.

4.3.2 Seed dispersal

Average recovery rate (% of tagged seeds recovered after 4 to 5 days) was 66% in 2014, 42% in 2015, and 46% in 2016. Of seeds that were recovered, in 2014, 36.5% were eaten without being moved away from the feeding site, 49.3% were moved and eaten, and 20.5% were moved and cached. In 2015 these numbers were 40.8%, 36.1%, and 23% respectively; in 2016 they were 66.1%, 29.0%, and 4.8% cached, respectively. In 2016 the apparent shift in proportions was driven by a low caching rate at the Indiana site rather than the addition of the Illinois site. Average dispersal distance for individual BC3s with more than one dispersal event ranged from 4.92 m to 9.08 m, which was less than the average for Chinese chestnut (10.49 m) and greater than the single American chestnut that was cached (1.85 m). Chinese chestnuts were dispersed farther and cached more frequently (29.6% of the time) than American chestnuts (3.2% of the time) and most BC3s (average over 14 families with at least one dispersed seed: 16.5%). Pools created for genotyping reflected the wide range of variation in dispersal. The average dispersal distance for seeds of trees placed in the strong-dispersal pool (seven BC3 and one Cm) was 9.08 ± 4.12 m and average caching frequency (no. cached / no. found) 25.3% (Table 1). For the moderate-dispersal pool (seven BC3) the average distance dispersed was 4.12 m ± 0.78 m, with a caching frequency of 8.2%. In the weak-dispersal pool (nine BC3 and one Cd), the average distance dispersed was 0.185± 0.59 m, and the caching frequency was 0.5%. Mean individual seed size was a statistically significant predictor of mean individual distance to caching in a simple linear regression where individuals with average seed dispersal distance 0 (i.e., seeds that were only recovered eaten at the feeding platform) were excluded ($t_{1, 25} = 4.43$, $p = 0.0002$, adjusted $r^2 = 0.42$) and seeds cached / total number of seeds recovered ($t_{1, 25} = 2.26$, $p = 0.03$, $r^2 = 0.14$) (Figure 1, Figure 2). In a binomial regression, mean seed mass was not a significant predictor of whether an individual mother tree had at least one seed recovered in a cache (z value = 1.394, p = 0.163).

4.3.3 Genotyping

Enough 100 bp paired-end reads (57-67 million) were obtained for each pool to cover the ~ 800 Mb chestnut genome between 7.2 and 8.5 times, so that each individual tree was represented by about one read at any locus in the genome. A small fraction of total bases (~2%) were removed from each sample by Trimmomatic due to low read quality prior to analysis. In the strong-dispersal pool, 341363 SNPs with coverage >8 were identified with one allele fixed in *Cm,* one in *Cd,* and a hybrid genotype in ‘Clapper’ (informative SNPs); 177884 were identified in the moderate-dispersal pool, and 215590 wer identified in in the weak-dispersal pool. ‘Clapper’ was hybrid at 50% of the loci in the genome that have one allele fixed for *Cm* and another for *Cd*.

4.3.4 Analysis of hybrid regions among pools

The mean value of the hybridity estimate (HE) over all SNPs in predicted genes with coverage ≥ 8 was highest for the strong-dispersal pool (0.44 ± 0.123) (Figure 4) and lowest for pool C (0.294 ± 0.164) (Figure 6). For the moderate-dispersal pool, the mean value of HE was 0.313 ± 0.174 over all predicted genes (Figure 5). When windows of 10 genes were used, the average difference in HE among windows was greatest between the high- and weak-dispersal pool (0.155 ± 0.088), but the difference between the strong- and moderate-dispersal pools was similar (0.137 ± 0.101) (Figure 7) and both were much larger than the average difference between strong- and moderate- dispersal pools (0.019 ± 0.085). Of 2714 bins of ten predicted genes, there was one region where the difference in HE between the strong-dispersal pool and the weak-dispersal pool was > 3 standard deviations greater than the mean difference, and 53 bins > 2 standard deviations above the mean. There were two bins for which the difference in HE between the strong-dispersal pool and the moderate-dispersal pool was > 3 SD above the mean, and 58 where the difference was > 2 SD above the mean. The heterozygosity estimate was a statistically significant predictor of Clapper heterozygosity (t value 6.12, p <0.0001) although the correlation coefficient was relatively low (r^2=0.123) (Figure 3).

The simulation-based method for determining the likelihood that more *Cm* alleles are present in a given pool than expected supported some loci (identified in bold, Table 3, Table 4) where the average percentile score indicated that the strong-dispersal pool's genotype at the gene was in the top 20% of the simulated distribution of *Cm* alleles in a random pool. Most of the candidate genes in Table 4 had at least one SNP located within the gene sequence with p <0.05 to reject the null hypothesis that variation in the number of *Cm* alleles in the strong-dispersal pool, but not the weak- or moderate-dispersal pools, was due to chance alone.

4.3.5 Annotations of genes within hybrid regions

Candidate genes for differences in seed dispersal were analyzed from 18 bins with the largest deviations from the mean difference in the heterozygosity estimate between the strong-dispersal pool and the moderate- and weak-dispersal pools.These 18 were chosen from an initial set of 28 candidate regions because their "Clapper" and American chestnut heterozygosity values unambiguously indicated that “Clapper” could have contributed a Cm allele to its BC3 descendants at these loci. Fourteen of these bins had a large difference in heterozygosity between the strong-dispersal pool (pool A) and the others (pools B and C); three were identified based on the difference between the strong- and moderate-dispersal pool; and one was identified based on the difference between strong- and weak-dispersal pools while the strong- and moderate-dispersal pools showed no difference (Table 2). Examining annotations of predicted genes in these regions revealed several that have plausible roles in seed development and subsequent seed handling and dispersal by squirrels (Table 3, Table 5). Many of the predicted genes in these regions aligned to cDNA sequences from chestnuts and other nut-bearing species in the order Fagales (Table 6, Table 7).

4.4 Discussion

4.4.1 Dispersal trials

Seed size, as measured by seed mass, was associated with both dispersal distance (Figure 1) and likelihood of caching vs. consumption (Figure 2), and is the most obvious difference between American and Chinese chestnut seeds. It was also highly variable among the backcrossed trees in the study, which were deliberately chosen for their extreme seed phenotypes (Table 1). The genetic differences underlying this phenotypic variation could involve genes that control the development and expansion of the seeds themselves, or seed size could be a secondary effect of genes that affect pollination and seed set, since larger chestnuts tend to be produced when fewer nuts are produced per burr, an example of a seed size/seed number tradeoff due to limited maternal resources (Venable 1992). Differences in seed number, however, tend to be more influenced by environment (nutrient availability, climate variables that affect pollination) so are somewhat unlikely candidates for heritable differences in seed size (Li and Li 2015) compared to traits that affect female floral parts directly. Genetically controlled maternal influences on seed size involve modifications to cell size, proliferation, and growth in the integument, endosperm, and embryo of a developing seed (Li and Li 2015).

Caching rates may have been somewhat low in this experiment because of the seed-tagging method, which put a hole in the shell and attached a conspicuous wire and flag to the seed (Xiao et al. 2006). In accordance with previous studies (Jansen et al. 2002, Xiao et al. 2005, Moore et al. 2007, Tamura et al. 2008) , seeds with greater mass tended to exhibit greater dispersal distance and greater likelihood of caching vs. immediate consumption by squirrels. Chinese chestnuts in this study were much larger than American chestnuts (Table 1), so BC3 hybrids with seeds more similar in dimensions to Chinese chestnut were more likely to be cached. Seed dispersal distance was frequently great enough (10 meters or more) to remove the seed from beneath the canopy of an average-sized mature chestnut in some individual BC3s, and it is likely that the farthest-dispersed seeds were not recovered due to distance from the feeding site; at the Eureka, IL site, untagged Chinese chestnuts were observed germinating > 50 m from a pair of isolated adult trees near the site. No seeds were observed to survive the winter in caches of tagged experimental seeds; they were all eaten by the end of seed trials in late December. The conspicuous flag may have made the seeds even easier for fox squirrels to recover than usual.

Observed caches of both tagged experimental seeds and untagged seeds from nearby chestnut trees at both sites indicated that burial was quite shallow. Typically, the seed would be pushed into a small depression (about equal to the depth of the seed) in the soil surface, not covered with soil, but rather with a thin layer of grass thatch and leaves. The author observed several non-tagged Chinese chestnuts cached at this shallow depth, with the top of the seed exposed, germinating in spring 2017 at the Eureka, IL site. The value of squirrel caches to seed survival has been demonstrated by numerous studies (Licthi et al. 2017). No viable chestnuts were left uncached or uneaten at either site over the winter, but the rapid desiccation (and viability loss) of chestnuts and acorns stored at cold temperatures is well-documented (LePrince et al. 1999, Iakovoglou 2010, Roach et al. 2010). So, it seems that the fox squirrel-chestnut relationship represents a true conditional mutualism: squirrels eat the vast majority of cached seeds, but the cache offers an improved site from germination over a site on the soil surface, i.e., the resting place of a seed passively dispersed by gravity from the parent tree.

4.4.2 Genotyping of BC3 pools

The hybridity estimate (HE) used to identify regions with a predominance of *Cm/Cd* genotypes in pools showed a relatively low correlation with observed heterozygosity in 'Clapper'. Part of the cause for this apparent inaccuracy could be the fact that 50% of the 'Clapper' genome is hybrid, while the expected portion of the genome that is hybrid in a BC3 is only 12.5%. Therefore, many genes that are highly heterozygous in 'Clapper' due to a *Cm/Cd* genotype would not be expected to be hybrid in BC3s. HE was intended to quantify the bias towards *Cm* alleles at a locus rather than heterozygosity per se. The HE statistic seems to have captured the elevated heterozygosity, relative to *Cd/Cd* loci, that is characteristic of *Cm/Cd* hybrid loci; 'Clapper' was much more heterozygous than *Cd* in most of the seed dispersal candidate regions (Table 3). Including Chinese chestnut in the strong-dispersal pool skewed the average heterozygosity estimate for that pool higher, while including American chestnut in the weak-dispersal pool seems to have skewed the heterozygosity estimate lower for that pool, which may have increased the likelihood of identifying spurious candidate regions; there were more *Cm* alleles in pool A at nearly every locus because there was a *Cm* individual in the sample. The simulation-based method for identifying loci with more *Cm* alleles than expected in the strong-dispersal pool, however, took differences in pool species composition into account when generating null distributions of *Cm* allele frequencies for each pool to avoid such spurious identifications of candidate loci. Most of the predicted genes identified as candidates using the HE method had at least one SNP with a p-value < 0.05.

4.4.3 Genomic loci associated with seed development

Several of the genomic loci identified could influence the wide range of variation in seed size in hybrid chestnut. The ubiquitin-protein ligase at the Sd05 locus has weak similarity (24/45 positive amino acid matches over a short stretch of a >400 residue peptide sequence from BLASTP) to a RING-type E3 ubiquitin-protein ligase that underlies the Gw2 grain weight locus in rice (Song et al. 2007). The SUPERMAN transcription factor and EMBRYONIC FLOWER 2-like (EMF) genes at loci Sd06 and Sd10, respectively, are more likely to directly influence seed size by regulation of development of female flower parts. EMF2 in *Arabidopsis* is a gene encoding a Polycomb group protein (Yoshida et al. 2001) that regulates vegetative growth and development by suppressing the flower-development program, specifically, by forming part of a protein complex that binds to chromatin and prevents expression of flowering-related genes. It is expressed during the early stages of seed development in *Arabidopsis* (Yoshida et al. 2001). The predicted EMF2 gene in chestnut only had strong transcript support from *C. mollissima* and *C. dentata*, and was one of 5 EMF2-like genes predicted in the entire chestnut genome. SUPERMAN is a zinc-finger transcription factor that functions to maintain boundaries among floral parts during development (Sakai et al. 1995) and the numbers of stamens and carpels in a flower (Gaiser et al. 1995); mutants display altered seed shapes. EMF2 in *Arabidopsis* represses the transcription of several floral homeotic genes (APETALA3 and PISTILLATA) that may also be negatively regulated by SUPERMAN. APETALA3's expression is restricted to developing flowers, while another APETALA gene , Ap2, has been associated with differences in seed mass (Jofuku et al. 2004). The Abr1-like transcription factor at Sd12, has a role in *Arabidopsis* related to seed germination, stress response, and repression of ABA-regulated genes. It is similar to APETALA2 in structure (Pandey et al. 2005) and may interact with the APETALA2 protein. The main candidate gene at locus Sd14 is an LATERAL ROOT PRIMORDIUM 1 homolog. As indicated by the name, this gene is involved in the initiation of lateral roots, but also has a role in floral development along with a number of similar proteins that act in a dosage-dependent manner (Kuusk et al. 2006). Loss-of-function mutations at this gene in *Arabidopsis*, in conjunction with the loss of similar regulatory genes, can cause malformed gynoecia. Differences in seed size between American and Chinese chestnut could be controlled by genes that act to regulate cell proliferation in the integument, which ultimately affects the final size of the seed; a cytochrome P450 oxidase gene in *Arabidopsis* was observed to affect seed size this way (Adamski et al. 2009). Alternatively, the mechanism could be modification of cell number and cell size in the embryo itself, as in APETALA2 mutants (Ohto et al. 2004).

Genes that control seed size by modifying development of the cotyledons were less likely to be identified by this experiment because only mothers were genotyped, so the paternal contribution to embryo development was not considered. The locus Sd04 may be directly or indirectly involved in the inheritance of seed size. Sd04 contains a set of predicted exocyst complex 5 component genes. The exocyst complexes in *Arabidopsis* have a variety of molecular roles related to cellular growth (Hala et al. 2008) in stems and pollen tubes, and they appear to be involved in the formation of the cell plate during cytokinesis (Fendrych et al.2010). Mutants at the SEC5a and SEC5b exocyst complex subunits in *Arabidopsis* show male infertility due to poor pollen tube growth (Hala et al. 2008). Poor pollen germination is not a particularly likely explanation for a seed dispersal-related locus because paternal effects were not tested in this study, but it is interesting that both subunits required for male fertility loss in *Arabidopsis* mutants are located at the same locus in chestnut. Transcriptomic data indicated that exocyst complex 5 component genes are highly conserved and, most likely, ubiquitously expressed in Fagales trees (Table 5, Table 6). If differences in these genes reduced pollen production rather than germination, more resources would be available for the tree to form seeds, which could lead to larger seeds on average and greater likelihood of caching. It is also possible that the exocyst genes at Sd04 represent pleiotropic loci that affect pollen production and female floral development.

4.4.4 The role of dormancy in caching decisions

In addition to seed size, squirrel caching behavior is influenced by the perceived dormancy status of the seed (Smallwood et al. 2001, Xiao 2009). Squirrels are less likely to cache a seed that is perceived as breaking dormancy or approaching germination (Moore et al. 2007). A large amount of variation in seed dispersal distance and caching likelihood among BC3 trees is unexplained by seed size (Figure 1, Figure 2) so other difference among BC3 seeds must be influencing the decisions of dispersers, and differences in dormancy (or the perception of dormancy) may be important. There is no documented difference in seed dormancy between American and Chinese chestnut—both species must undergo a dormant period of several months to germinate (Saielli et al. 2012), and naturally begin germination in late winter and early spring, so it is unlikely that any seeds in the dispersal trials were approaching germination. European chestnuts may have a shorter dormancy period, germinating in winter (Baskin and Baskin 1998), but if this were true of Chinese chestnut it would most likely lead to trees with a Chinese allele at dispersal-related loci being dispersed a shorter distance, assuming that a shorter dormancy period leads to earlier germination. However, hybrid phenotypes are not always intermediate between parents (Woeste et al. 1998), so gene interactions could cause unpredictable phenotypes in interspecific hybrids. Nuts that resulted from a hybrid pollination with Allegheny chinkapin (*Castanea pumila*) as pollen parent and *Cm* as the seed parent exhibited reduced seed dormancy (Jaynes 1963, Metaxas 2013).

Since chestnuts are recalcitrant seeds, the seed is metabolically active during its dormant phase (LePrince et al. 1999, Roach et al. 2010). Sugar content of chestnuts under cold storage increases while starches diminish (Ertan et al. 2015). As these changes occur inside the seed, waxes on the pericarp gradually diminish and low molecular-weight compounds from the kernel begin to penetrate the pericarp (Sundaram 2016). Volatile compounds may be an important olfactory cue for rodent seed predators, with the waterproof wax layers on the pericarp masking seed volatile compounds (Paulsen et al. 2013). Therefore, seeds with a thick layer of wax on the pericarp might be perceived as more dormant than a seed with relatively thin layer of pericarp wax; Steele et al. (2001) demonstrated that a germinating white oak embryo inside a “dormant” red oak shell is perceived as dormant by squirrels.

4.4.5 Genomic loci involved in dormancy and the perception of dormancy

Several genes in seed dispersal-associated intervals appear to have a role in lipid metabolism that may be related to the formation and/or degradation of pericarp wax layers. Non-specific lipid transfer proteins (Sd15) in *Arabidopsis* are involved in the formation of suberin in crown galls (Deeken et al. 2016), various tissues of tomato in response to drought stress (Trevino et al. 1998), and the surface wax of broccoli leaves (Pyee et al. 1994). A cytochrome P450 oxidase (Sd11) similar to an *Arabidopsis* gene that is involved in fatty acid biosynthesis (Benveniste et al. 1998, Pinot and Beisson 2011) could have a direct role in suberin formation. A longevity assurance homolog 1-like gene (Sd12) is also likely to be involved in fatty acid biosynthesis, specifically, synthesis of ceramides (Ternes et al. 2011). Ceramides function in programmed cell death and signalling in plants (Liang et al. 2003) so this locus may be involved in stress response or hormone signalling (Markham et al. 2011) rather than accumulation of surface waxes.

The importance of fatty-acid genes in regulating squirrel dispersal was explored by Sundaram (2016), who showed that differences in the outer wax layer of the pericarp influence squirrel perception of seed dormancy. A gene that could affect the production of volatile compounds is the 2-alkenal reductase-like gene at Sd01, which is similar to a gene from tobacco (Mano et al. 2005) that detoxifies lipid peroxide-derived reactive carbonyls. Different alleles of this gene in BC3 chestnuts could affect the rate at which lipids are modified in the lead-up to germination. A gene potentially related to a somewhat different aspect of seed metabolism and chemical changes preceding germination is the alpha-amylase like gene predicted at Sd12; alpha-amylase is a key enzyme in breaking down storage starches in germinating seeds (Huang et al. 1992); the most similar gene to the predicted chestnut gene is one from mung bean that is expressed at a high level during germination (Tripathi et al. 2007).

In addition to the molecular signals from within the seed, which are transmitted through the pericarp and interpreted by squirrels as cues of germination, the pericarp itself is a regulator of seed dormancy. In peach (Martinez-Gomez and Dicenta 2001) and almond (Garcia-Gusano et al. 2004) dormancy can be reduced by removing the pericarp. The seed coat provides a physical barrier to germination in many plant species (Finch-Savage and Leubner-Metzger 2006) such that longer dormancy is conferred by a thicker seed coat. Since the pericarp and testa are composed entirely of maternal tissue, and our methods (grouping seeds by mother tree and genotyping only the mother tree) could only detect maternal genetic effects on seeds, loci involving the synthesis of the pericarp are particularly relevant. A number of the loci identified as candidates based on a higher number of *Cm/Cd* genotypes in the strong-dispersal pool versus the moderate- and weak-dispersal pools contain some type of cellulose synthase (Sd06, Sd16), pectin-modifying enzyme (Sd08, Sd17) or gene otherwise involved in formation of the cell wall (Sd03, Sd07), and these could alter germination schedules and squirrel behavior if they alter the formation of the pericarp.

A seed with a less permeable pericarp could be inferred to release fewer volatile compounds and absorb less water, delaying germination and reducing the likelihood of seed consumption (Paulsen et al. 2014). Sundaram et al. (2015) compared shell thickness of Chinese, American, and BC3 chestnuts and found that American chestnut had the thickest shell of the three and BC3 the thinnest. Shell thickness in BC3 chestnut was not linearly related to predicted genomic content of the parent species, *Cm* and *Cd.* Seed moisture content in the same study was highest in *Cm* and lowest in *Cd*, with BC3 intermediate. The pericarp anatomy of oaks has been studied more thoroughly than that of chestnuts, but the makeup of oak and chestnut seeds is quite similar; outside the embryo is a thin seed coat or testa surrounded by a pericarp with three layers, a thin inner layer, a parenchymatous middle layer, and the lignified outermost layer, the exocarp (Bonner and Vozzo 1987). In a study of *Quercus robur* seeds, Nikolic et al. (2010) found considerable variation among genotypes in pericarp thickness, with the thickness of the lignified exocarp negatively correlated with the thickness of the parenchyatous mesocarp. A thicker mesocarp may be associated with chestnut oak (*Quercus montana*)'s ability to germinate in dry soils (Korstian 1927, McQuilkin 1990) by increasing its ability to absorb and retain water, while a thicker lignified endocarp in cork oak (*Quercus suber*) serves as protection against excess water loss, and also could inhibit germination by increasing the mechanical strength of the pericarp (Sobrino-Vesperinas and Viviani 2000). This information is interesting because American chestnut was formerly a common associate of chestnut oak on dry ridges in the Appalachian Mountains (Wang et al. 2013), and might have a similar adaptive syndrome in its seeds. If a water-absorbent pericarp with a thinner exocarp and thicker mesocarp layer is advantageous in this habitat, it might also occur in American chestnut. A thicker, but more permeable, pericarp could contribute to earlier germination of American chestnuts than Chinese chestnuts and influence squirrels' caching decisions. We found that American chestnuts lost more mass, in cold storage than Chinese chestnuts presumably due to drying, indicating that the pericarp of American chestnut may be more permeable; alternatively, the greater percent of mass lost in *Cd* could be the result of the surface area/volume ratio of smaller seeds. Although BC3 families were highly variable, those that were more likely to be cached rather than eaten lost less mass over the same storage period than American chestnut. A thick, but permeable pericarp could be an adaptiation to absorb any available water and speed up germination in the relatively dry soils of American chestnut's preferred Appalachian habitat. If Chinese chestnut lacks this adaptation and BC3 trees display a phenotype similar to Chinese chestnut rather than American chestnut, it may hamper the establishment of restored hybrid populations in the Appalachians. Despite over a century of widespread planting in the eastern United States, Chinese chestnut has only been observed to naturalize extremely rarely (Miller et al. 2014). This is probably in large part due to poor competition with native species, which can easily overtop Chinese chestnut, but poor seedling establishment may have been overlooked as a contributing factor.

As the pericarp matures in some fleshy fruits, cellulose synthase genes are expressed during the ripening process (e.g. Shangguan et al. 2017) due to the remodeling of cell walls. Some cellulose synthase genes are essential for cell wall formation (Li et al. 2009), so they could be important early in seed development in altering the size and shape of the seed or the thickness of the pericarp. Predicted cellulose synthase genes were found at several seed dispersal candidate loci, one at Sd06 (CSLG2-like) and another at Sd16 (CESA2-like). For the Sd06 cellulose synthase, the most similar gene in *Arabidopsis* is a membrane-localized beta-glycan synthase expressed primarily in young seedlings (Richmond and Somerville 2001) so it may be involved in cellular remodeling as the seed approaches germination. This cellulose synthase appears to fairly conserved within *Castanea* but not among all Fagales trees, based on transcriptome data. If differences in these cellulose synthase genes affect the rate at which the seed breaks dormancy and begins to emit volatiles that squirrels associate with germination, it could affect caching behavior. The most similar *Arabidopsis* gene to the Sd17 cellulose synthase, CESA2, has been implicated in embryo development (Beeckman et al. 2002) but is ubiquitously expressed, and mutants show decreased growth and seed production due to abnormal cell expansion (Chu et al. 2007). This gene could be involved in the expansion of embryonic cells leading up to germination, or to the elongation of cells during seed formation, or both. Transcripts aligned to this predicted protein from every Fagales tree tested. A UDP-glucuronic acid decarboxylase 2-like predicted gene (Sd07) is also likely to contribute to the synthesis of cell wall polysaccharides; it is a membrane-bound protein involved in the formation of xyloglucan (Harper and Bar-Peled 2002). It is not possible to tell whether these predicted genes might be involved in the initial formation of cell walls in the pericarp, or in their remodeling during the germination process. The UXS2-like gene was only supported by transcripts from the extensive oak cDNA libraries, which indicates that it may have a tissue-specific function and was not captured in any available chestnut mRNA-sequencing data.

The two pectinesterase-like predicted genes (Sd08 and Sd17) are similar to genes in *Arabidopsis* that are expressed in developing siliques (seed pods) (Louvet et al. 2006). Pectin methlyesterase genes are active in the woody tissues of poplar as young wood tissue, which has high pectin content, transitions to mature woody tissue, which is mostly lignin with some remaining pectin (Mellerowicz et al. 2001). The pectinesterase genes at this locus could be involved in the formation of the lignified pericarp. Pectinesterase genes, however, are also active within seeds during the transition from dormancy to germination in yellow cedar (*Chamaecyparis nootkaensis*) (Ren and Kermode 2000). Specifically, investigators found that PME activity gradually increased throughout cold storage and peaked during germination and hypothesized that the pectin methylesterases broke down cell walls in the endosperm immediately surrounding the embryo prior to radical emergence. *Arabidopsis* lines with overexpression of a PME inhibitor showed more rapid germination, indicating a role for PME in regulating dormancy and germination (Müller et al. 2013). The pectinesterase at the Sd08 locus is expressed in several chestnut species and most of the Fagales transcriptomes examined, but the one at the Sd17 locus was only found in oak transcriptomes (Table 5, Table 6). The oak transcriptomes examined were more extensive in terms of tissue samples, so this gene may only be locally expressed in seed; the other may be more ubiquitously expressed.

At Sd03, a leucine-rich repeat extensin like gene (similar to Arabidopsis LRX3) was predicted. This predicted gene has strong support from transcripts across the Fagales; its Arabidopsis homolog belongs to a vegetative clade of the LRR-extensin like gene family (Baumberger et al. 2003); it is involved in the formation of cell walls and is expressed throughout the plant. The exocyst complex component genes at Sd04, discussed above in relation to pollen production, could also affect the cellular composition of the pericarp given their role in cell division and polarity. An oligosaccharide synthesis gene similar to galactinol synthase 2 (GOLS2) in *Arabidopsis* was predicted at the Sd06 locus. The *Arabidopsis* homolog of this gene is involved in generating raffinose family oligosaccharides, which are thought to protect seeds from desiccation and are associated with seed maturation in soybeans and maize (Castillo et al.

1990, Taji et al. 2002). Furthermore, galactinol and raffinose synthesis increases protection against oxidative damage due to osmotic stress caused by cold or salinity (Nishizawa et al. 2008), and oxidative conditions inside seeds are believed to be what lead to the death of desiccated chestnut embryos (Roach et al. 2010). Desiccated (dead) chestnuts are rapidly overgrown with microbes internally (Roach et al. 2010), which would destroy their food value for squirrels. Squirrels are sensitive to weevil damage (Steele et al. 1996), caching damaged seeds less frequently. If squirrels are also less likely to cache desiccated, nonviable seeds, a seed with greater resistance to desiccation would be more attractive for caching.

Plant hormones play an important role in the transition from dormant to germinated seed; abscisic acid ABA is associated with the maintenance of seed dormancy while gibberellins are associated with the switch to germination (Rodriguez-Gacio et al. 2009). A negative regulator of ABA signaling in beech (*Fagus sylvatica*) is believed to contribute to the switch from dormancy to germination in beechnuts (Gonzalez-Garcia et al. 2003). Ethylene also functions to counteract the effects of ABA and promote seed germination (Corbineau et al. 2014), and salycilic acid may act in conjunction with ABA to suppress the gibberellins-promoted germination program (Xie et al. 2007). Genes at several of the seed-dispersal candidate loci described here appear to be involved in hormone-signalling pathways that affect germination. At locus Sd12, there is a predicted ethylene-responsive transcription factor similar to the ABR1 locus in *Arabidopsis*, which is involved in repressing ABA signaling during the transition to seed germination. It also has a role in stress response, which may be the reason there were transcripts similar to the predicted gene found in most of the examined Fagales transcriptomes. At Sd02, there is a predicted abscisic acid stress-ripening protein, which is similar to a gene in tomato (Kalifa et al. 2004) that is associated with fruit ripening under stressful conditions. Other ASR genes have some involvement in sugar trafficking and carbohydrate metabolism (Golan et al. 2014). If the chestnut gene here is involved in maintaining or promoting dormancy through the influence of ABA, its Chinese chestnut allele could affect dispersal by delaying the germination-associated changes in seed chemistry that are detected by squirrels. The last locus with an apparent role in hormone signaling is the DLO2-like gene at Sd13. The *Arabidopsis* homolog for this predicted gene breaks down salicylic acid (Zeilmaker et al. 2015). In Xie et al. (2007) salicylic acid was observed to suppress alpha-amylase activity associated with seed germination. It is possible that both the predicted alpha-amylase at Sd12 is suppressed by salicylic acid, which is broken down by the DLO2-like enzyme coded at Sd13, and that the Chinese chestnut alleles for these genes confers a slower transition from dormancy to germination in the BC3 seeds that were more likely to be cached and carried farther before caching.

Squirrels perceive volatile compounds from seeds as cues of metabolic activity and impending germination (Sundaram 2016); these volatile compounds are thought to escape the pericarp as it becomes more porous and germination approaches. One particularly interesting locus appeared to contain a cluster of four volatile terpene synthase genes, which are most similar to terpene synthesis genes highly expressed in the fruits of strawberry (Aharoni et al. 2004); these genes are thought to influence flavor and aroma profiles of the ripening strawberry fruit. Nerolidol is a sesquiterpene compound found in many plants (Chan et al. 2016). Sundaram (2016) found the release of beta-amyrin, a triterpene, to be associated with germination of chestnuts. While these compounds are distantly related, their synthesis may be metabolically linked by production of the intermediate squalene. Nerolidol synthase converts farnesyl diphosphate (FPP) to nerolidol. In a yeast study, overexpressing FPP synthase and squalene synthase greatly increased beta-amyrin production (Zhang et al. 2015). Beta-amyrin has been associated with wax degradation in other plants (Buschhaus and Jetter 2012), so it could degrade the cuticular waxes of the outer pericarp as it is released, preparing the seed for germination (Sundaram 2016). The genes found here do not directly influence beta-amyrin, but could influence upstream production of substrate molecules or divert carbon away from beta-amyrin production. Interestingly, of the three nerolidol synthase-like genes at this locus, one showed evidence of expression in Chinese chestnut and two others showed evidence of expression in American chestnut and oaks, but not Chinese chestnut (Table 5), possibly indicating interspecific differences in the expression of these genes. None of the nerolidol synthase genes were expressed in the root libraries of Japanese and European chestnut, and there was little evidence of their expression in the non animal-dispersed taxa examined (*Alnus, Betula*) nor in *Fagus*. If the expression of multiple copies of nerolidol synthase in American chestnut leads to an increase in the activity of volatile organic compounds that degrade pericarp waxes, it could lead BC3 seeds that express the *Cd* alleles to germinate sooner, and cause squirrels to eat rather than cache these seeds.

4.5 Conclusions

Our dispersal trials supported the results of Blythe et al. (2015) showing that BC3 chestnut seeds have a dispersal phenotype (likelihood of caching and distance to cache) similar to, but distinct from, American chestnut, and extended them by documenting that Chinese chestnuts are more likely to be cached and carried a long distance than American and BC3 chestnut seeds. We were also able to demonstrate that, as predicted by the wide range of variability in seed size of individual BC3 mother trees, individual BC3s have variable dispersal phenotypes. Some extreme individuals displayed a similar dispersal phenotype to Chinese chestnut, but most are similar to American chestnut.

Results from whole-genome sequencing of strong-dispersal (long distance, large percentage of seeds dispersed) low-dispersal (short distance, small percentage of seeds dispersed) and near-zero dispersal (nearly all seeds eaten on the spot) revealed regions that carried a Chinese allele and an American allele in the strong-dispersal pool and two American chestnut alleles in one or both low-dispersal pools. Heterozygosity values from whole-genome sequences of ‘Clapper,’ the BC1 resistance donor for all BC3 trees in the study, and from a pair of American chestnuts confirmed that the strong-dispersal group could have inherited a *Cm* allele at these seed dispersal candidate loci.

Predicted genes in the putative dispersal loci include some that could plausibly alter the structure of female inflorescences, namely, genes with similarity to the floral homeotic genes SUPERMAN and EMBRYONIC FLOWER 2 of *Arabidopsis thaliana.* Many of the loci contained predicted genes that apparently pertained to cell wall modification, including pectin methylesterases and cellulose synthases, some of which were similar to *Arabidopsis* genes that are most highly expressed in developing seeds and siliques and others that are expressed in germinating seeds. These genes are most likely related to seed dormancy and the perception of dormancy by squirrels, either modulating the thickness and chemical makeup of the seed pericarp or changes that occur within the seed as the transition from dormancy to germination begins. Other genes related to seed dormancy and the perception of seed dormancy by squirrels included lipid synthesis and modification enzymes that likely play a role in the formation of the waxy coat of the outer pericarp, volatile terpenes that may be involved in the breakdown of that waxy layer and olfactory perceptions of the loss of dormancy, enzymes that modulate hormone signalling shifts that occur as dormancy ends (suppression of ABA and SA and promotion of ethylene signalling), and an alpha-amylase that is a candidate for converting starches to sugars during the germination process. Many of these candidate genes aligned to cDNA sequences from species in *Castanea*, *Quercus*, and other animal-dispersed and non-animal-dispersed taxa in the Fagales.

Taken as a whole, this work supports the hypothesis that germination cues are a major factor in the caching decisions of squirrels. The putative gene loci identified here could be important in the evolutionary history of Fagales and their coevolution with conditional mutualist seed dispersers. More research is needed to determine the exact nature of differences in seed dormancy between American and Chinese chestnut, and variation in the coding sequences of predicted genes among different chestnut species and their relatives in the Fagaceae.

4.6 Literature cited

Adamski NM, Anastasiou E, Erksson S, O'Neill CM, Lenhard M (2009) Local maternal control of seed size by KLUH/CYP78A5-dependent growth signalling. Proceedings of the National Academy of Sciences USA 106(47):20115-20.

Baker, H.G. 1972. Seed weight in relation to environmental conditions in California. Ecology 53:997-1010.

Aharoni A, Giri AP, Verstappen FW, Bertea CM, Sevenier R, Sun Z, Jongsma MA, Schwab W, Bouwmeester HJ (2004) Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species. Plant Cell 16:3110-3131.

Baskin CC, Baskin JM (1998) Seeds: Ecology, Biogeography, and Evolution of Dormancy and Germination. Elsevier, 666 pp.

Baumberger N, Doesseger B, Guyot R, Diet A, Parsons RL, Clark MA, Simmons MP, Bedinger P, Goff SA, Ringli C, Keller B (2003) Whole-genome comparison of leucine-rich repeat extensions in *Arabidopsis* and rice. A conserved family of cell wall proteins form a vegetative and reproductive clade. Plant Physiology 131(3):1313-1326.

Beeckman T, Przemeck GK, Stamatiou G, Lau R, Terryn N, De Rycke R, Inzé D, Berleth T (2002) Genetic complexity of *cellulose synthase a* gene function in *Arabidopsis* embryogenesis. Plant Physiology 130(4):1883-93.

Benveniste I, Tijet N, Adas F, Philipps G, Salaun JP, Durst F (1998) CYP86A1 from Arabidopsis thaliana encodes a cytochrome P450-dependent fatty acid omega-hydroxylase. Biochemical and Biophysical Research Communications 243(3):688-93.

Blythe RM, Lichti NI, Smyser TJ, Swihart RK (2015) Selection, caching and consumption of hardwood seeds by forest rodents: implications for restoration of American chestnut. Restoration Ecology 23(4):473-481.

Bonner FT, Vozzo JA (1987) Seed biology and technology of *Quercus*. USDA Forest Service Southern Forest Expt Stn General Tech Rept 50-66. New Orleans, Louisiana.

Brodin, A. 2010. The history of scatter hoarding studies. Philosophical Transactions of the Royal Society of Biology 365: 869-881.

Buschhaus C, Jetter R (2012) Composition and physiological function of the wax layers coating *Arabidopsis* leaves: B-Amyrin negatively affects the intracuticular water barrier. Plant Physiology 160(2):1120-1129.

Calhane, V.H. 1942. Caching and recovery of food by the western fox squirrel. The Journal of Wildlife Management 6(4):338-352.

Castillo EM, De Lumen BO, Reyes PS, De Lumen HZ (1990) Raffinose synthase and galactinol synthase in developing seeds and leaves of legumes. Journal of Agricultural and Food Chemistry 38:351-355.

Chan W-K, Tan LT-H, Chan K-G, Lee L-H, Goh B-H (2016) Nerolidol: a sesquiterpene alcohol with multi-faceted pharmacological and biological activities. Molecules 21(5):529.

Chu Z, Chen H, Zhang Y, Zhang Z, Zheng N, Yin B, Yan H, Zhu L, Zhao X, Yuan M, Zhang X, Xie Q (2007) Knockout of the AtCESA2 gene affects microtubule orientation and causes abnormal cell expansion in *Arabidopsis*. Plant Physiology 143(1):213-24.

Clarke, M.F., Kramer, D.L. 1994. Scatter-hoarding by a larder-hoarding rodent: intraspecific variation in the hoarding behavior of the eastern chipmunk, *Tamias striatus*. Animal Behavior 48:299-308.

Connor, K., Donahoo, J., Schafer, G. 2006. How does prolonged exposure to natural conditions affect acorn moisture and viability? Connor, Kristina F., ed. 2006. Proceedings of the 13th biennial southern silvicultural research conference. Gen. Tech. Rep. SRS–92. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 640 p.

Corbineau F, Xia Q, Bailly C, El-Maarouf-Bouteau H (2014) Ethylene, a key factor in the regulation of seed dormancy. Frontiers in Plant Science 5:539.

Dalgleish, HJ, Swihart RK (2012) American chestnut past and future: implications of restoration for resource pulses and consumer populations of eastern U.S. forests. Restoration Ecology 20(4): 490-497.

Deeken R, Saupe S, Klinkenberg J, Riedel M, Leide J, Hedrich R, Mueller TD (2016) The nonspecific lipid transfer protein AtLtpl-4 is involved in suberin formation of *Arabidopsis thaliana* crown galls. Plant Physiol 172(3):1911-1927.

Ertan E, Erdal E, Gulsum A, Algul BE (2015) Effects of different postharvest storage methods on the quality parameters of chestnuts (*Castanea sativa* Mill.) HortScience 50(4):577-581.

Farrant JM, Pammenter NW, Berjak P (1989) Germination-associated events and the desiccation sensitivity of recalcitrant seeds – a study on three unrelated species. Planta 178:189-198.

Finch-Savage WE, Leubner-Metzger G (2006) Seed dormancy and the control of germination. New Phytologist 171(3):501-523.

Fox, J.F. 1982. Adaptation of gray squirrel behavior to autumn germination by white oak acorns. Evolution 36(4):800-809.

Friis, E.M., Crane, P.R., Pederson, K.R. 2011. Early Flowers and Angiosperm Evolution. Cambridge University Press, ISBN 978-0-521-59283-3.

Fry SC, Aldington S, Hetherington PR, Aitken J (1993) Oligosaccharides as signals and substrates in the plant cell wall. Plant Physiology 103:1-5.

Gaiser JC, Robinson-Beers K, Gasser CS (1995) The *Arabidopsis* SUPERMAN gene mediates asymmetric growth of the outer integument of ovules. The Plant Cell 7:333-345.

Goheen, J.R., and R.K. Swihart. 2003. Food-hoarding behavior of gray squirrels and North American red squirrels in the central hardwoods region: implications for forest regeneration. Canadian Journal of Zoology 81:1636-1639.

Golan I, Dominguez PG, Konrad Z, Shkolnik-Inbar D, Carrari F, et al. (2014) Tomato *ABSCISIC STRESS RIPENING (ASR)* gene family revisited. PloS One 9(10):e107117. doi:10.1371/journal.pone.0107117.

Gomez JM (2004) Bigger is not always better: conflicting selective pressures on seed size in *Quercus ilex*. Evolution 58(1):71-80.

Gonzalez-Garcia MP, Rodriguez D, Nicolas C, Rodriguez PL, Nicolas G, Lorenzo O (2003) Negative regulation of abscisic acid signaling by the *Fagus sylvatica* FsPP2C1 plays a role in seed dormancy regulation and promotion of seed germination. Plant Physiology 133(1):135-144.

Gutierrez R, Lindeboom JJ, Paredez AR, Emons AMC, Ehrhardt DW (2009) *Arabidopsis* cortical microtubules position cellulose synthase delivery to the plasma membrane and interact with cellulose synthase trafficking components. Nature Cell Biology 11:797-806.

Garcia-Gusano M, Martinez-Gomez P, Dicenta F (2004) Breaking seed dormancy in almond (*Prunus dulcis* (Mill.) D.A. Webb) Scientia Horticulturae 99(3):363-370.

Hirsch, B.T., Kays, R., Jansen, P.A. 2012. A telemetric thread tag for tracking seed dispersal by scatter-hoarding rodents. Plant Ecology 213:933-943.

Howe HF, Smallwood J (1982) Ecology of seed dispersal. Annual Review of Ecology and Systematics 13:201-228.

Huang N, Stebbins GL, Rodriguez RL (1992) Classification and evolution of α-amylase genes in plants. Proc Natl Acad Sci USA (1992) 89:7526-7530.

Iakovoglou V, Misra MK, Hall RB, Knapp AD (2010) Alterations of seed variables under storage in nitrous oxide (N2O) atmospheres for two recalcitrant *Quercus* species. Scandinavian Journal of Forest Research 25:24-30.

Ivan, J.S., Swihart, R.K. 2000. Selection of mast by granivorous rodents of the central hardwood forest region. Journal of Mammalogy 81(2):549-562.

Jansen, P. A., Bartholomeus, M., Bongers, F., Elzinga, J. A., den Ouden, J., & Van Wieren, S. E. (2002). 14 The Role of Seed Size in Dispersal by a Scatter-hoarding Rodent. Seed Dispersal and Frugivory: Ecology, Evolution, and Conservation, 209.Janzen, D.H. 1971. Seed predation by animals. Annual Review of Ecology and Systematics 2:465-492.

Jacobs, D.F. 2007. Toward development of silvical strategies for forest restoration of American chestnut (*Castanea dentata*) using blight-resistant hybrids. Biological Conservation 137(4):497-506.

Jacobs, D.F., Dalgleish, H.J., Nelson, C.D. 2012. A conceptual framework for restoration of threatened plants: the effective model of American chestnut (*Castanea dentata*) reintroduction. New Phytologist 197: 378-393.

Johnson, W.C., Webb, T. III. 1989. The role of blue jays (*Cyanocitta cristata* L.) in the postglacial dispersal of fagaceous trees in eastern North America. Journal of Biogeography 16(6):561-571.

Jofuku KD, Omidyar PK, Gee Z, Okamuro JK (2004) Control of seed mass and seed yield by the floral homeotic gene APETALA2. Proc Nat Acad Sci USA 102(8):3117-3122.

Korstian CF (1927) *Factors controlling germination and early survival in oaks.* Bulletin No. 19, Yale University School of Forestry, New Haven, CT, 115 p.

Kuusk S, Sohlberg JJ, Magnus ED, Sundberg E (2006) Functionally redundant SHI family genes regulate *Arabidopsis* gynoecium development in a dose-dependent manner. The Plant Journal 47:99-111.

Larios E, Burquez A, Becerra JX, Venable DL (2014) Natural selection on seed size through the life cycle of a desert annual plant. Ecology 95(11):3213-3220.

Larson-Johnson K (2015) Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. New Phytologist 209(1):418-435.

Leprince O, Buitnik J, Hoekstra FA (1999) Axes and cotyledons of recalcitrant seeds of *Castanea sativa* Mill. exhibit contrasting responses of respiration to drying in relation to desiccation sensitivity. Journal of Experimental Botany 50(338):1515-1524.

Leopold DJ, McComb WC, Muller RN (1998) Trees of the Central Hardwood Forests of North America: An Identification and Cultivation Guide. Timber Press, Portland, Oregon, 509 p. ISBN-13: 978-0881924060.

Lesur I, Bechade A, Lalanne C, Klopp C, Noirot C, Leple J-C, Kremer A, Plomion C, Le Provost G (2015) A unigene set for European beech (*Fagus sylvatica* L.) and its use to decipher the molecular mechanisms involved in dormancy regulation. Molecular Ecology Resources 15(5):1192-1204.

Lesur I, Le Provost G, Bento D, et al. (2015) The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release. BMC Genomics 16:112.

Li M, Xiong G, Li R, Cui J, Tang D, Zhang B, Pauly M, Cheng Z, Zhou Y (2009) Rice cellulose synthase-like D4 is essential for normal cell-wall biosynthesis and plant growth. The Plant Journal 60(6):1055-1069.

Li N, Li Y (2015) Maternal control of seed size in plants. Journal of Experimental Botany 66(4):1087-1097.

Liang H, Yao N, Song JT, Luo S, Lu H, Greenberg JT (2003) Ceramides modulate programmed cell death in plants. Genes and Development 17:2636-2641.

Lichti NI, Steele MA, Swihart RK (2017) Seed fate and decision-making processes in scatter-hoarding rodents. Biological Reviews 92(1):474-504

Lohbeck M, Lebrija-Trejos E, Martinez-Ramos M, Meave JA, Poorter L, Bongers F (2015) Functional trait strategies of trees in dry and wet tropical forests are similar but differ in their consequences for succession. PLoS One https://doi.org/10.1371/journal.pone.0123741

Louvet R, Cavel E, Gutierrez L, Guenin S, Roger D, Gillet F, Guerineau F, Pelloux J (2006) Comprehensive expression profiling of the pectin methylesterase gene family during silique development in *Arabidopsis thaliana*. Plant 224:782-791.

Mano J, Belles-Boix E, Babiychuk E, et al. (2005) Protection against photooxidative injury of tobacco leaves by 2-alkenal reductase. Detoxification of lipid peroxide-derived reactive carbonyls. Plant Physiology 139(4):1773-1783.

Markham JE, Molino D, Gissot L, Bellec Y, Hematy K, Marion J, Belcram K, Palauqui J-C, Satiat-JeuneMaitre B, Faure J-D (2011) Sphingolipids containing very-long-chain fatty acids define a secretory pathway for specific polar plasma membrane protein targeting in *Arabidopsis.* The Plant Cell 23:2362-2378.

Martinez-Garcia PJ, Crepeau MW, Puiu D et al. (2016) The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. The Plant Journal 87(5):507-32.

Martinez-Gomez P, Dicenta F (2001) Mechanisms of dormancy in seeds of peach (Prunus persica (L.) Batsch) cv. GF305. Scientia Horticulturae 91(1):51-58.

McQuilkin RA (1990) *Quercus prinus* L. Chestnut Oak. In. Burns RM, Honkala BH (tech. coords.) Silvics of North American, Vol. 2: Hardwoods, Agriculture Handbook 654, USDA Forest Service, Washington, D.C., 726 p.

Mellerowicz EJ, Baucher M, Sundberg B, Boerjan W (2001) Unravelling cell wall formation in the woody dicot stem. Plant Molecular Biology 47:239-274.

Mendu V, Griffiths JS, Persson S, Stork J, Downie AB, Voinicuc C, Haughn GW, DeBolt S (2011) Subfunctionalization of cellulose synthases in seed coat epidermal cells mediates secondary radial wall synthesis and mucilage attachment. Plant Physiology 157:441-453.

Miller AC, Woeste KE, Anagnostakis SL, Jacobs DF (2014) Exploration of a rare population of Chinese chestnut in North American: stand dynamics, health and genetic relationships. AoB Plants 6:plu065.

Moore, J.E., McEuen, A.B., Swihart, R.K., Contreras, T.A., Steele, M.A. 2007. Determinants of seed removal distance by scatter-hoarding rodents in deciduous forests. Ecology 88(10):2529-2540.

Mu H-Z, Liu Z-J, Lin L, Li H-Y, Jiang J, Liu G-F (2012) Transcriptomic analysis of phenotypic changes in birch (*Betula platyphylla*) autotetraploids. International Journal Molecular Science 13:13012-13029.

Müller K, Levesque-Tremblay G, Bartels S, Weitbrecht K, Wormit A, Usadel B, Haughn G, Kermode AR (2005) Demethylesterification of cell wall pectins in *Arabidopsis* plays a role in seed germination. Plant Physiology 161:305-316.

Nikolic NP, Merkulov LS, Krstic BD, Pajevic SP, Borisev MK, Orlovic SS (2010) Variability of acorn anatomical characteristics in *Quercus robur* L. genotypes. Proc. Nat. Sci, Matica Srpska Novi Sad 118:47-58.

Nishizawa A, Yabuta Y, Shigeoka S (2008) Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. Plant Physiology 147(3): 1251-63.

Ohto MA, Fisher RL, Goldberg RB, Nakamura K, Harada JJ (2004) Control of seed mass by APETALA2. Proceedings of the National Academy of Sciences USA 102(8):3123-3128.

Pandey GK, Grant JJ, Cheong YH, Kim BG, Li L, Luan S (2005) ABR1, an APETALA2-domain transcription factor that functions as a repressor of ABA response in *Arabidopsis.* Plant Physiology 139(3):1185-93.

Paulsen TR, Hogstedt G, Thompson K, Vandvik V, Eliassen S, Leishman M (2014) Conditions favoring hard seededness as a dispersal and predator escape strategy. Journal of Ecology 102(6):1475-1484.

Pinot F, Beisson F (2011) Cytochrome P450 metabolizing fatty acids in plants: characterization and physiological roles. FEBS Journal 278(2):195-205.

Pollard M, Beisson F, Li Y, Ohlrogge (2008) Building lipid barriers: biosynthesis of cutin and suberin. Trends in Plant Science 13(5):236-246.

Pyee J, Yu HS, Kolattukudy PE (1994) Identification of a lipid transfer protein as the major protein in the surface wax of broccoli (*Brassica oleracea*) leaves. Archives of Biochemistry and Biophysics 311(2):460-468.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ren C, Kermode AR (2000) An increase in pectin methyl esterase activity accompanies dormancy breakage and germination of yellow cedar seeds. Plant Physiology 124(1):231-242.

Richardson KB, Licthi NI, Swihart RK (2013) Acorn-foraging preferences of four species of free-ranging avian seed predators in eastern deciduous forests. The Condor 115(4): 863-873.

Richmond TA, Somerville CR (2001) Integrative approaches to determining Csl function. Plant Molecular Biology 47:131-143.

Roach T, Beckett RP, Minibayeva FV, Colville L, Whitaker C, Chen H, Bailly C, Kranner I (2009) Extracellular superoxide production, viability, and redox poise in response to desiccation in recalcitrant *Castanea sativa* seeds. Plant, Cell & Environment 33:59-75.

Robbins CT, Mole S, Hagerman AE, Hanley TA (1987) Role of tannins in defending plants against ruminants: reduction in dry matter digestion? Ecology 68(6): 1606-1615.

Rodriguez-Gacio M, Matilla-Vazquez MA, Matilla AJ (2009) Seed dormancy and ABA signalling: the breakthrough goes on. Plant Signalling and Behavior. 4(11):1035-1048.

Rom S, Gilad A, Kalifa Y, Konrad Z, Karpasas MM, Goldgur Y, Bar-Zvi D (2006) Mapping the DNA- and zinc-binding domains of ASR1 (abscisic acid stress ripening), an abiotic-stress regulated plant specific protein. Biochimie 88:621-628.

Rowley ER, Fox SA, Bryant DW, Sullivan C, Givan S, Mehlenbacher SA, Mockler TA (2012) Assembly and characterization of the European hazelnut (*Corylus avellana* L.) 'Jefferson' transcriptome. Crop Science 52:2679-2686.

Rutter PA, Miller G, Payne JA (1991) Chestnuts (*Castanea*). ISHS Acta Horticulturae 290: Genetic Resources of Temperate Fruit and Nut Crops (761-788).

Saielli TM, Schaberg PG, Hawley GJ, Halman JM, Gurney KM (2012) Nut cold hardiness as a factor influencing the restoration of American chestnut in northern latitudes and high elevations. Canadian Journal of Forest Research 42(5):859-857.

Sakai H, Medrano LJ, Meyerowitz EM (1995) Role of SUPERMAN in maintaining *Arabidopsis* floral whorls. Nature 378:199-203.

Schlink K (2009) Identification and characterization of differentially expressed genes from *Fagus sylvatica* roots after infection with *Phytophthora citricola*. Plant Cell Reports 28(5):873-82.

Schupp EW, Jordano P, Gomez JM (2010) Seed dispersal effectiveness revisited: a conceptual review. New Phytologist 188: 333-353.

Senter SD (1994) Comparison of total lipids, fatty acids, sugars and nonvolatile organic acids in nuts from four *Castanea* species. Journal of the Science of Food and Agriculture 65:223-227.

Shangguan L, Mu Q, Fang X, Zhang K, Jia H, Li X, Bao Y, Fang J (2017) RNA-sequencing reveals biological networks during table grapevine ('Fujiminori') fruit development. PLoS One journal.pone.0170571.

Sobrino-Vesperinas E, Viviani AB (2000) Pericarp micromorphology and dehydration characteristics of *Quercus suber* acorns. Seed Sci Res 10:401-407.

Smallwood PD, Steele MA, Faeth SH (2001) The ultimate basis of the caching preferences of rodents, and the oak-dispersal syndrome: tannins, insects, and seed germination. American Zoologist 41:840-851.

Song X-J, Huang W, Shi M, Zhu M-Z, Lin H-X (2007) A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin protein ligase. Nature Genetics 39:629-630.

Stapanian MA, Smith CC (1978) A model for seed scatterhoarding: coevolution of fox squirrels and black walnuts. Ecology 59:884-896.

Staton ME, Addo-Quaye C, Cannon N, Tomsho LP, Drautz D, Wagner TK, Zembower N, Ficklin S, Saski C, Burhans R, Schuster SC, Abbott AG, Nelson CD, Hebard FV, Carlson JE (2014) *The Chinese chestnut (Castanea mollissima) genome version 1.1*, http://www.hardwoodgenomics.org/chinese-chestnut-genome, Access date August 2, 2016.

Steele MA, Contreras TA, Hadj-Chikh LZ, Agosta SJ, Smallwood PD, Tomlinson CN (2014) Behavioral Ecology 25(1):206-215.

Steele MA, Hadj-Chikh LZ, Hazeltine J (1996) Caching and feeding decisions by *Sciurus carolinensis*: responses to weevil-infested acorns. J Mammology 77(2):305-314.

Steele MA, Smallwood PD, Spunar A, Nelsen E (2001) The proximate basis of the oak dispersal syndrome: detection of seed dormancy by rodents. American Zoologist 41:852-864.

Sundaram M, Willoughby JR, Licthi NI, Steele MA, Swihart RK (2015) Segregating the effects of seed traits and common ancestry of hardwood trees on eastern gray squirrel foraging decisions. PLoS One https://doi.org/10.1371/journal.pone.0130942.

Sundaram M (2016) The role of seed attributes in eastern gray squirrel foraging. Ph.D. Disseration, Purdue University.

Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. The Plant Journal 29(4):417-426.

Tamura N (2001) Walnut hoarding by the Japanese wood mouse, *Apodemus speciosus* Temminck. Journal of Forest Research 6:187-190.

Tamura N, Hashimoto Y, Hayashi F (1999) Optimal distances for squirrels to transport and hoard walnuts. Animal Behaviour 58:635-642.

Tamura N, Hayashi F (2008) Geographic variation in walnut seed size correlates with hoarding behavior of two rodent species. Ecological Research 23:607-614.

Ternes P, Feussner K, Werner S, Lerche J, Iven T, Heilmann I, Riezman H, Feussner I (2011) Disruption of the ceramide synthase LOH1 causes spontaneous cell death in *Arabidopsis thaliana*. New Phytologist 192(4):841-54.

Theimer TC (2005) Chapter 17. Rodent scatterhoarders as conditional mutualists. In: Seed Fate, eds. P.M. Forget, J.E. Lambert, P.E. Hulme, S.B. Vander Wall, CAB international, p. 283.

Thompson DC, Thompson PS (1980) Food habits and caching behavior of urban grey squirrels. 58(5): 701-710.

Thomson FJ, Moles AT, Auld TD, Kingsford RT (2011) Seed dispersal distance is more strongly correlated with plant height than with seed mass. Journal of Ecology 99(6):1299-1307.

Thorington RW, Koprowski JL, Steele MA, Whatton JF (2012) <u>Squirrels of the World.</u> Johns Hopkins University Press.

Tompsett PB, Pritchard HW (1998) The effect of chilling and moisture status on the germination, desiccation tolerance and longevity of *Aesculus hippocastanum* L. seed. Ann Bot 82(2):249-261.

Torales S, Rivarola M, Pomponio MF et al. (2012) Transcriptome survey of Patagonian southern beech *Nothofagus nervosa* (= *N. alpina*): assembly, annotation, and molecular marker discovery. BMC Genomics 13(1):291. DOI: 10.1186/1471-2164-13-291

Trevino MB, O'Connell MA (1998) Three drought-responsive members of the nonspecific lipid-transfer protein gene family in *Lycopersicon pennellii* show different developmental patterns of expression. Plant Physiology 116(4): 1461-68.

Tripathi P, Lo Leggio L, Mansfield J, Ulbrich-Hofmann R, Kayastha AM (2007) Alpha-amylase from mung beans (*Vigna radiata*)—correlation of biochemical properties and tertiary structure by homology modeling. Phytochemistry 68(12):1623-31.

Ueno S, Taguchi Y, Tomaru N, Tsumura Y (2009) Development of EST-SSR markers from an inner bark cDNA library of *Fagus crenata.* Conservation Genetics 10:1477.

Vander Wall SB, Thayer TC, Hodge JS, Beck MJ, Roth JK (2001) Scatter-hoarding behavior of deer mice (*Peromyscus maniculatus*). Western North American Naturalist 61(1):109-113.

Vander Wall SB (2001) The evolutionary ecology of nut dispersal. Botanical Review 67(1): 74-117.

Vander Wall SB, Jenkins SH (2003) Reciprocal pilferage and the evolution of food-hoarding behavior. Behavioral Ecology 14(5):656-667.

Vander Wall SB, Kuhn KM, Beck MJ (2005) Seed removal, seed predation, and secondary dispersal. Ecology 86(3): 801-806.

Venable DL (1992) Size-number trade-offs and the variation of seed size with plant resource status. The American Naturalist 140(2):287-304.

Wang B, Chen J (2012) Effects of fat and protein levels on foraging preferences of tannin in scatter-hoarding rodents. PLOS One 7(7):e40640.

Wang GG, Knapp BO, Clark SL, Mudder BT (2013) The silvics of *Castanea dentata* (Marsh.) Borkh., American chestnut, Fagaceae (Beech Family). Gen. Tech. Rep. SRS-GTR-173. Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station, 18 p.

Worthen LM, Woeste KE, Michler CH (2010) Breeding American chestnuts for blight resistance. Plant Breeding Reviews 33: 305-339.

Xiao Z, Jansen PA, Zhang Z (2006) Using seed-tagging methods for assessing post-dispersal seed fate in rodent-dispersed trees. Forest Ecology and Management 223:18-23.

Xiao Z, Gao X, Jiang M, Zhang Z (2009) Behavioral adaptation of Pallas's squirrels to germination schedule and tannins in acorns. Behavioral Ecology doi: 10.1093/beheco/arp096.

Xiao Z, Gao X, Steele MA, Zhang Z (2009) Frequency-dependent selection by tree squirrels: adaptive escape of nondormant white oaks. Behavioral Ecology 21:169-175. doi: 10.1093/beheco/arp169.

Xiao Z, Gao X, Zhang S (2013) Sensitivity to seed germination schedule by scatter-hoarding Pere David's rock squirrels during mast and non-mast years. Ethology 119(6): 472-479.

Xiao S, Zhang Z, Wang Y (2005) Effects of seeds size on dispersal distance in five rodent-dispersed fagaceous species. Acta Oecologica 28:221-229.

Xie Z, Zhang Z-L, Hanzlik S, Cook E, Shen QJ (2007) Salycilic acid inhibits gibberellins-induced alpha-amylase expression and seed germination via a pathway involving an abscisic-acid-inducible WRKY gene. Plant Mol Biol 64(3):293-303.

Yalifa Y, Gilad A, Konrad Z, Zaccai M, Scolnik PA, Bar-Zvi D (2004) The water- and salt-stress regulated *Asr1* (abscisic acid stress ripening) gene encodes a zinc-dependent DNA-binding protein. Biochemical Journal 381(2):373-378.

Yoshida N, Yanai Y, Chen L, Kato Y, Hiratsuka J, Miwa T, Sung ZR, Takahashi S (2001) EMBRYONIC FLOWER2, a novel polycomb group protein homolog, mediates shoot development and flowering in *Arabidopsis*. The Plant Cell 13:2471-2481.

Zhang G, Cao Q, Liu J, Liu B, Li J, Li C (2015) Refactoring beta-amyrin synthesis in *Saccharomyces cerevisiae.* AIChE 61(10):3172-3179.

Zwolak R, Crone EE (2012) Quantifying the outcome of plant-granivore interactions. Oikos 121:20-27.

Table 4.1. Summary of seed dispersal data for trees in three genotyping pools, showing the pool each individual parent tree was assigned to, its mean seed weight, the number of seeds cached, the average distance seeds were cached away from the feeding site, and the total number of seeds found for that individual (cached + eaten).

Pool	Species	Year	Mean seed mass (g)	N (cached)	Mean distance (m)	Total found	Total offered
A[a]	CC	2014-15	8.53	8	10.49	27	75
A	BC1	2015	3.77	5	8.25	20	25
A	BC1	2015	6.82	13	7.341	21	25
A	BC1	2016	4.14	1	10.98	15	30
A	BC1	2016	3.35	2	7.8	13	30
A	BC1	2014	4.51	7	9.08	19	25
A	BC1	2014	4.03	2	9.64	28	50
A	BC1	2014	2.55	6	6.40	30	50
B[b]	BC1	2014	3.55	3	4.53	23	50
B	BC1	2016	3.16	2	4.92	20	40
B	BC1	2016	4.57	1	4.42	25	50
B	BC1	2016	2.67	1	3.69	23	40
B	BC1	2016	3.50	1	2.85	9	30
B	BC1	2016	3.82	1	3.52	13	30
B	BC1	2016	3.48	1	4.92	9	30
C[c]	BC1	2016	2.53	0	0	12	20
C	BC1	2016	3.22	0	0	9	40
C	BC1	2016	4.01	0	0	24	50
C	BC1	2016	3.66	0	0	14	20
C	BC1	2014	2.99	0	0	24	50
C	BC1	2014	3.47	0	0	24	50
C	BC1	2015	2.54	0	0	19	25
C	BC1	2015	1.47	0	0	7	25
C	BC1	2014	3.09	0	0	24	50
C	AC	2014-15	1.52	1	1.85	30	50

[a]Strong-dispersal pool; [b]Moderate-dispersal-pool; [c]Weak-dispersal pool

Table 4.2. Summary of regions associated with differences in seed dispersal among BC3 chestnuts and chestnut species, including pseudochromosome (LG) and putative function of the locus based on protein alignments to predicted polypeptides.

Locus	LG	BP interval	Pools[a]	St. dev[b]	Predicted genes
Sd1	LGA	37471441-37753244	A:BC	>2	Reduction of double bonds in enones
Sd02	LGA	67866962-67971352	A:BC	>2	Ripening, response to abscisic acid
Sd03	LGA	83772158-84089674	A:BC	>3	Cell wall synthesis, flower development
Sd04	LGA	106688319-106815059	A:BC	>2	Acceptance of pollen
Sd05	LGB	8460333-8666294	A:BC	>2	Oligosaccharide synthesis, response to dehydration
Sd06	LGB	27019449-27249144	A:BC	>2	Hemicellulose, cell wall synthesis
Sd07	LGB	41287540-41436283	A:BC	>3	Polysaccharide synthesis
Sd08	LGC	24974261-25133355	A:BC	>2	Cell wall modification
Sd09	LGC	29530300-29997579	A:BC	>2	Lipid metabolism
Sd10	LGC	48273510-48654688	A:C	>2	Flower development
Sd11	LGE	21212476-21654063	A:BC	>2	Fatty acid biosynthesis
Sd12	LGG	33993229-34224973	A:B	>2	Hormone signaling related to seed germination
Sd13	LGG	40408926-40687727	A:B	>2	Defense against fungi, tissue senescence
Sd14	LGH	46084023-46409469	A:BC	>2	Vascular development of female floral parts
Sd15	LGI	25672061-25978423	A:BC	>2	Lipid transport
Sd16	LGL	58234900-58423908	A:BC	>2	Primary cell wall formation
Sd17	LGL	59955058-60179331	A:BC	>2	Cell wall modification
Sd18	LGL	65537114-65794683	A:B	>2	Volatile terpene synthesis

[a] Shows pattern of divergence among pools that led to identification; A:BC = likely hybrid in strong-diserpsal pool (A) but not in moderate (B) or weak-dispersal (C) pools, A:B = likely hybrid in A but not B, A:C = likely hybrid in A but not C. [b] Standard deviations in excess of the mean difference in hybridity estimator (HE) between pools A:BC, A:B, or A:C for the given interval.

Table 4.3. Selected candidate genes with summary of evidence for involvement in seed dispersal. Boldface indicates candidate genes with the strongest evidence

Site	Gene	Pool	Pool	Pool	Het.[d]	Het.[e]	UniProt[g]	%
		A[a]	B[b]	C[c]	BC1	*Cd*		ID
Sd01	A.g4714	0.488	0.026	0.067	0.120	0.013	DBR_TOBAC	64
Sd02	A.g8635	0.750	0.088	0.242	0.176	0.062	ASR1_SOLLC	82
Sd03	**A.g10648**	**0.650**	**0.145**	**0.206**	**0.742**	**0.000**	**LRX3_ARATH**	**77**
Sd03	**A.g10657**	**0.522**	**0.077**	**0.195**	**0.522**	**0.042**	**VIL1_ARATH**	**48**
Sd04	A.g13357	0.585	0.124	0.154	0.026	0.055	SEC5A_ARATH	57
Sd04	A.g13359	0.471	0.104	0.063	0.007	0.043	SEC5B_ARATH	71
Sd05	B.g1118	0.470	0.093	0.082	0.108	0.035	GOLS2_ARATH	72
Sd06	B.g3452	0.422	0.089	0.119	0.119	0.021	SUP_ARATH	41
Sd06	B.g3458	0.534	0.031	0.180	0.305	0.088	CSLG2_ARATH	36
Sd06	B.g3460	0.522	0.000	0.174	0.227	0.068	CSLG2_ARATH	38
Sd07	B.g5164	0.449	0.289	0.260	0.459	0.183	UXS2_ARATH	70
Sd08	C.g3074	0.437	0.077	0.093	0.236	0.042	PME31_ARATH	79
Sd09	C.g3725	0.432	0.399	0.062	0.755	0.008	AAPT1_ARATH	85
Sd09	C.g3728	0.516	0.287	0.061	0.787	0.025	AAPT1_ARATH	98
Sd10	C. g6050	0.583	0.325	0.155	0.838	0.068	EMF2_ARATH	46
Sd12	G.g4366	0.175	0.000	0.483	0.857	0.033	ABR1_ARATH	55
Sd12	G.g4369	0.484	0.012	0.369	0.55	0.130	AMYA_VIGMU	71
Sd12	G.g4370	0.460	0.018	0.434	0.592	0.198	LAG12_ARATH	66
Sd13	**G.g5214**	**0.557**	**0.066**	**0.580**	**0.446**	**0.175**	**DLO2_ARATH**	**39**
Sd14	H.g5907	0.647	0.078	0.072	0.093	0.018	LRP1_ARATH	50
Sd15	**I. g3291**	**0.730**	**0.000**	**0.070**	**0.429**	**0.000**	**NLTL5_ARATH**	**33**
Sd15	I.g3304	0.550	0.000	0.082	0.600	0.000	C94A2_VICSA	52
Sd16	L.g7302	0.535	0.066	0.108	0.203	0.092	CESA2_ARATH	78
Sd17	**L.g7556**	**0.517**	**0.050**	**0.129**	**0.160**	**0.031**	**PME51_ARATH**	**59**
Sd18	L.g8191	0.516	0.236	0.419	0.671	0.155	NES1_FRAAN	61
Sd18	**L.g8192**	**0.597**	**0.112**	**0.316**	**0.700**	**0.022**	**TPS13_RICCO**	**62**
Sd18	L.g8198	0.667	0.098	0.429	0.554	0.142	NES1_FRAVE	58
Sd18	L.g8208	0.527	0.078	0.455	0.667	0.123	NES2_FRAAN	60

[a] Pool of individuals with highest mean dispersal distance and largest % of seeds cached rather than consumed; [b] Pool of individuals with lower mean dispersal distance and lower

caching %; [c]Pool of seeds that were rarely or never cached; [d]Proportion of heterozygous snps in gene for 'Clapper' calculated using Vcftools; [e]Proportion of heterozygous snps in gene for *Cd* calculated in Vcftools; [f] Percentile of expected *Cm* allele frequency distribution based on 1,000,000 simulated pooled genotypes for each pool, averaged over all SNP loci within gene.

Table 4.4. Statistical evidence for *Cm* allele abundance at candidate loci in BC3 chestnuts with the highest rates of dispersal success.

Site	Gene	UniProt[a]	A[b]	B[c]	C[d]	nSnps, pool A[e]	nSnps <0.05, pool A[f]
Sd01	A.g4714	DBR_TOBAC	0.35	0.91	0.87	11	0
Sd02	A.g8635	ASR1_SOLLC	na	na	na	0	0
Sd03	**A.g10648**	**LRX3_ARATH**	**0.17**	**0.59**	**0.59**	**10**	**2**
Sd03	**A.g10657**	**VIL1_ARATH**	**0.19**	**0.87**	**0.41**	**19**	**8**
Sd04	A.g13357	SEC5A_ARATH	na	na	na	0	0
Sd04	A.g13359	SEC5B_ARATH	0.46	0.91	0.88	2	1
Sd05	B.g1118	GOLS2_ARATH	0.49	0.91	na	8	0
Sd06	B.g3452	SUP_ARATH	0.57	na	0.88	2	0
Sd06	B.g3458	CSLG2_ARATH	0.24	0.91	0.36	5	0
Sd06	B.g3460	CSLG2_ARATH	0.72	0.91	na	5	0
Sd07	B.g5164	UXS2_ARATH	na	na	na	0	0
Sd08	C.g3074	PME31_ARATH	0.39	0.91	0.77	6	1
Sd09	C.g3725	AAPT1_ARATH	0.48	0.35	0.82	91	6
Sd09	C.g3728	AAPT1_ARATH	0.44	0.38	0.76	47	4
Sd10	C. g6050	EMF2_ARATH	0.24	0.37	0.82	26	8
Sd12	G.g4366	ABR1_ARATH	na	na	0.17	0	0
Sd12	G.g4369	AMYA_VIGMU	0.45	0.89	0.22	38	4
Sd12	G.g4370	LAG12_ARATH	0.28	na	na	2	0
Sd13	**G.g5214**	**DLO2_ARATH**	**0.06**	**0.91**	**0.03**	**6**	**4**
Sd14	H.g5907	LRP1_ARATH	0.38	0.84	0.82	9	2
Sd15	**I. g3291**	**NLTL5_ARATH**	**0.05**	**0.00**	**na**	**1**	**0**
Sd15	I.g3304	C94A2_VICSA	0.64	0.91	0.61	1	0
Sd16	L.g7302	CESA2_ARATH	0.48	0.91	0.86	10	0
Sd17	**L.g7556**	**PME51_ARATH**	**0.05**	**0.91**	**na**	**2**	**0**
Sd18	L.g8191	NES1_FRAAN	0.43	0.91	0.35	3	1
Sd18	**L.g8192**	**TPS13_RICCO**	**0.18**	**0.91**	**0.42**	**50**	**10**
Sd18	L.g8198	NES1_FRAVE	0.39	0.91	0.27	31	3
Sd18	L.g8208	NES2_FRAAN	0.30	0.91	0.42	9	2

[a]Closest UniProt homolog; [b]p-value for random distribution of *Cm* alleles in the strong-dispersal pool averaged over all SNPs in the gene; [c]p-value for random distribution of *Cm*

alleles in the moderate-dispersal pool averaged over all SNPs in the gene; [d]p-value for random distribution of *Cm* alleles in the weak-dispersal pool averaged over all SNPs in the gene; [e]number of informative SNP loci within gene (informative = different alleles fixed in *Cm* and *Cd*, hybrid in ‘Clapper’); [f]number of informative SNPs with $p < 0.05$ for over-representation of *Cm* alleles in pool A.

Table 4.5. Descriptions of the best aligned proteins for predicted chestnut genes identified in regions of the chestnut genome that have *Cm* and *Cd* alleles in frequently dispersed trees and only *Cd* alleles in less-dispersed trees.

Locus	Gene	UniProt	Full name	Description
Sd01	A.g4714	DBR_TOBAC	2-alkenal reductase	Modifies many organic compounds by reducing C=C double bonds
Sd02	A.g8635	ASR2_SOLLC	Abscisic stress-ripening 2	Modulate expression of sugar-regulated genes
Sd03	A.g10648	LRX3_ARATH	Leucine-rich extension-like 3	Regulates cell wall formation
Sd03	A.g10657	VIL1_ARATH	VIN3-like protein 1	Promotes short-day flowering, vernalization
Sd04	A.g13357	SEC5A_ARATH	Exocyst complex component	Involved in primary cell wall formation
Sd04	A.g13359	SEC5B_ARATH	Exocyst complex component	Involved in primary cell wall formation
Sd05	B.g1111	UPL6_ARATH	Ubiquitin protein ligase	Ubiquination and degradation of target proteins
Sd05	B.g1118	GOLS2_ARATH	Galactinol synthase 2	Synthesis of osmoprotecant oligosaccharides
Sd06	B.g3452	SUP_ARATH	SUPERMAN transcript regulator	Regulates floral development
Sd06	B.g3458	CSLG2_ARATH	Cellulose synthase-like protein G2	Beta-glycan synthase; polymerizes cell wall hemicelluloses
Sd07	B.g5164	UXS2_ARATH	UDP-glucoronic acid decarboxylase 2	Carbohydrate biosynthesis; glycosaminoglycan biosynthesis
Sd08	C.g3074	PME31_ARATH	Pectinesterase 31	Results in cell wall rigidification
Sd09	C.g3725	AAPT1_ARATH	Choline/ethanolaminephosphotransferase 1	Phospholipid metabolism
Sd10	C. g6050	EMF2_ARATH	EMBRYONIC FLOWER 2	Polycomb group protein; regulator of flower development
Sd11	E.g2720	C86A1_ARATH	Cytochrome P450 86A1	Suberin biosynthetic process
Sd12	G.g4366	ABR1_ARATH	Ethylene-responsive TF	May be involved in seed germination via abscisic acid pathway

Table 4.5 continued

Locus	Gene	UniProt	Full name	Description
Sd12	G.g4369	AMYA_VIGMU	Alpha-amylase	Starch hydrolysis
Sd12	G.g4370	LAG12_ARATH	Longevity assurance homolog 2	Fatty acid biosynthesis
Sd13	G.g5214	DLO2_ARATH	DMR6-like oxygenase 2	Salicylic acid catabolic process
Sd14	H.g5907	LRP1_ARATH	LATERAL ROOT PRIMORDIUM 1	Involved in formation and vasculature of female flower parts
Sd15	I. g3291	NLTL5_ARATH	Non-specific lipid transfer	Binding, transport of lipids
Sd15	I.g3304	C94A2_VICSA	Cytochrome P450 94A2	Hydroxylation of fatty acids
Sd16	L.g7302	CESA2_ARATH	Cellulose synthase	Beta-1,4-glucan microfibril crystallization, cell wall formation
Sd17	L.g7556	PME51_ARATH	Pectinesterase	Demethylesterification of cell wall pectin
Sd18	L.g8191	NES1_FRAAN	Nerolidol synthase	Monoterpene and sesquiterpene biosynthesis; expressed in fruit
Sd18	L.g8192	TPS13_RICCO	Terpene synthase	Sesquiterpene synthase
Sd18	L.g8208	NES2_FRAAN	Nerolidol synthase	Monoterpene and sesquiterpene biosynthesis

Table 4.6. Candidate genes from putative hybrid regions in highly-dispersed chestnuts with summary of alignments to publicly-available transcriptome sequences from trees in the Fagaceae and Nothofagaceae showing percent sequence identity for the best cDNA from each species.

Locus	Gene	UniProt	*Cm*[a]	*Cd*[b]	*Cc*[c]	*Cs*[d]	*Qa*[e]	*Qp*[f]	*Qr*[g]	*Fg*[h]	*Fs*[i]	*Nn*[j]
Sd01	A.g4714	DBR	100	-	100	97.6	95.0	97.7	99.4	-	-	80.8
Sd02	A.g8635	ASR2	-	-	-	-	-	65.4	-	-	-	-
Sd03	A.g10648	LRX3	100	100	97.9	98.2	93.1	95.6	95.1	95.5	-	-
Sd03	A.g10657	VIL1	100	100	100	93.1	96.1	96.2	94.9	-	-	83.4
Sd04	A.g13357	SEC5A	86.3	99.3	100	100	97.0	97.1	96.8	89.6	-	-
Sd04	A.g13359	SEC5B	100.0	98.6	99.1	98.7	90.0	97.1	-	-	-	-
Sd05	B.g1111	UPL6	100.0	100.0	99.2	95.7	95.0	92.6	91.6	87.5	96.1	98.2
Sd05	B.g1118	GOLS2	-	-	-	-	-	-	-	-	-	-
Sd06	B.g3452	SUP	-	-	-	-	-	-	-	-	-	-
Sd06	B.g3458	CSLG2	100.0	100.0	95.4	-	-	95.8	87.0	87.2	82.9	66.2
Sd06	B.g3460	CSLG2	-	-	-	-	-	-	-	-	-	-
Sd07	B.g5164	UXS2	-	-	-	-	-	85.2	100	-	-	-
Sd08	C.g3074	PME31	100	99.2	-	-	96.6	95.3	96.2	-	-	71.7
Sd09	C.g3725	AAPT1	81.5	95.5	81.0	90.7	87.8	90.3	87.9	-	81.9	-
Sd09	C.g3728	AAPT1	89.1	-	-	100	-	100	100	-	91.3	-
Sd10	C. g6050	EMF2	98.9	98.4	-	-	-	-	-	-	-	-
Sd11	E.g2720	C86A1	-	74.6	-	-	-	-	-	-	81	-
Sd12	G.g4366	ABR1	67.7	86.6	73.4	74.1	49.5	64.4	81.5	-	-	-
Sd12	G.g4369	AMYA	99.3	98.8	-	-	-	97.0	97.0	87.7	89.9	-
Sd12	G.g4370	LAG12	98.9	88.5	-	98.9	94.4	95.5	97.7	-	-	83.0
Sd13	G.g5214	DLO2	100	99.4	-	-	-	96.4	72.5	-	-	-
Sd14	H.g5907	LRP1	79.4	100	80.2	69.6	100	60.5	88.9	-	-	-
Sd15	I. g3291	NLTL5	89.0	100.0	85.0	-	-	70.7	77.3	-	-	-
Sd15	I.g3304	C94A2	-	-	-	-	95.9	-	-	-	-	-
Sd16	L.g7302	CESA2	85.6	100.0	90.0	88.3	100.0	100.0	87.3	86.0	96.7	67.8
Sd17	L.g7556	PME51	-	-	-	-	-	93.5	85.0	-	-	-
Sd18	L.g8191	NES1	93.1	-	-	-	-	-	-	-	-	-
Sd18	L.g8192	TPS13	-	-	-	-	100.0	-	-	-	-	-
Sd18	L.g8198	NES1	-	92.1	-	-	-	100	95.7	-	-	-
Sd18	L.g8208	NES2	-	100	-	-	-	95.5	90.8	-	-	-

[a]*Cm = Castanea mollissima,* [b]*Cd = Castanea dentata,* [c]*Cc = Castanea crenata,* [d]*Cs = Castanea sativa,* [e]*Qa = Quercus alba,* [f]*Qp = Quercus petraea/robur,* [g]*Qr = Quercus rubra,* [h]*Fg = Fagus grandifolia,* [i]*Fs = Fagus sylvatica,* [j]*Nn = Nothofagus nervosa*

Table 4.7. Candidate genes from putative hybrid regions in highly-dispersed chestnuts with summary of alignments to transcriptome sequences from trees in theJuglandaceae and Betulaceae showing percent sequence identity for the best cDNA from each species.

Locus	Gene	UniProt	*Cm*[a]	*Jr*[b]	*Jn*[c]	*Ca*[d]	*Aru*[e]	*Arh*[f]	*Bet*[g]
Sd01	A.g4714	DBR	100	84.2	82.2	85.8	82.5	82.1	85.0
Sd02	A.g8635	ASR2	-	-	-	44.2	41.8	41.8	68.0
Sd03	A.g10648	LRX3	100	-	94.3	95.7	94.3	92.5	90.3
Sd03	A.g10657	VIL1	100	83.0	83.0	84.4	92.4	89.0	87.9
Sd04	A.g13357	SEC5A	86.3	86.4	86.7	-	46.2	93.2	88.5
Sd04	A.g13359	SEC5B	100.0	-	-	91.9	91.5	93.3	-
Sd05	B.g1111	UPL6	100.0	80.5	82.0	98.0	94.0	96.2	85.8
Sd05	B.g1118	GOLS2	-	84.0	-	-	-	-	-
Sd06	B.g3452	SUP	-	60.0	-	-	-	-	-
Sd06	B.g3458	CSLG2	100.0	74.7	73.3	66.0	84.2	74.5	81.1
Sd06	B.g3460	CSLG2	-	-	-	94.6	-	-	-
Sd07	B.g5164	UXS2	-	25.7	87.5	-	-	89.2	86.7
Sd08	C.g3074	PME31	100	87.9	89.2	88.6	88.4	85.2	89.2
Sd09	C.g3725	AAPT1	81.5	80.5	79.7	86.3	87.1	87.9	87.1
Sd09	C.g3728	AAPT1	89.1	-	87.0	-	84.8	-	-
Sd10	C. g6050	EMF2	98.9	-	-	-	-	-	74.3
Sd11	E.g2720	C86A1	-	89.8	91.2	96.1	-	-	-
Sd12	G.g4366	ABR1	67.7	40.6	-	-	-	-	47.2
Sd12	G.g4369	AMYA	99.3	-	-	85.2	71.7	-	73.2
Sd12	G.g4370	LAG12	98.9	84.1	84.1	85.2	66.7	-	84.1
Sd13	G.g5214	DLO2	100	-	-	-	-	-	-
Sd14	H.g5907	LRP1	79.4	58.3	63.0	-	-	-	82.6
Sd15	I. g3291	NLTL5	89.0	50.3	65.1	61.8	-	-	59.8
Sd15	I.g3304	C94A2	-	-	-	-	-	-	-
Sd16	L.g7302	CESA2	85.6	84.4	88.5	100	79.1	60.5	81.9
Sd17	L.g7556	PME51	-	-	-	-	-	-	-
Sd18	L.g8191	NES1	93.1	-	38.9	-	-	-	-
Sd18	L.g8192	TPS13	-	-	-	-	-	-	-
Sd18	L.g8198	NES1	-	-	-	-	-	-	-
Sd18	L.g8208	NES2	-	68.2	81.2	81.2	-	-	78.8

[a]*Cm = Castanea mollissima*, [b]*Jr = Juglans regia*, [c]*Jn = Juglans nigra*, [d]*Ca = Corylus avellana*, [e]*Aru = Alnus rubra*, [f]*Arh = Alnus rhombifolia*, [g]*Bet = Betula* spp.

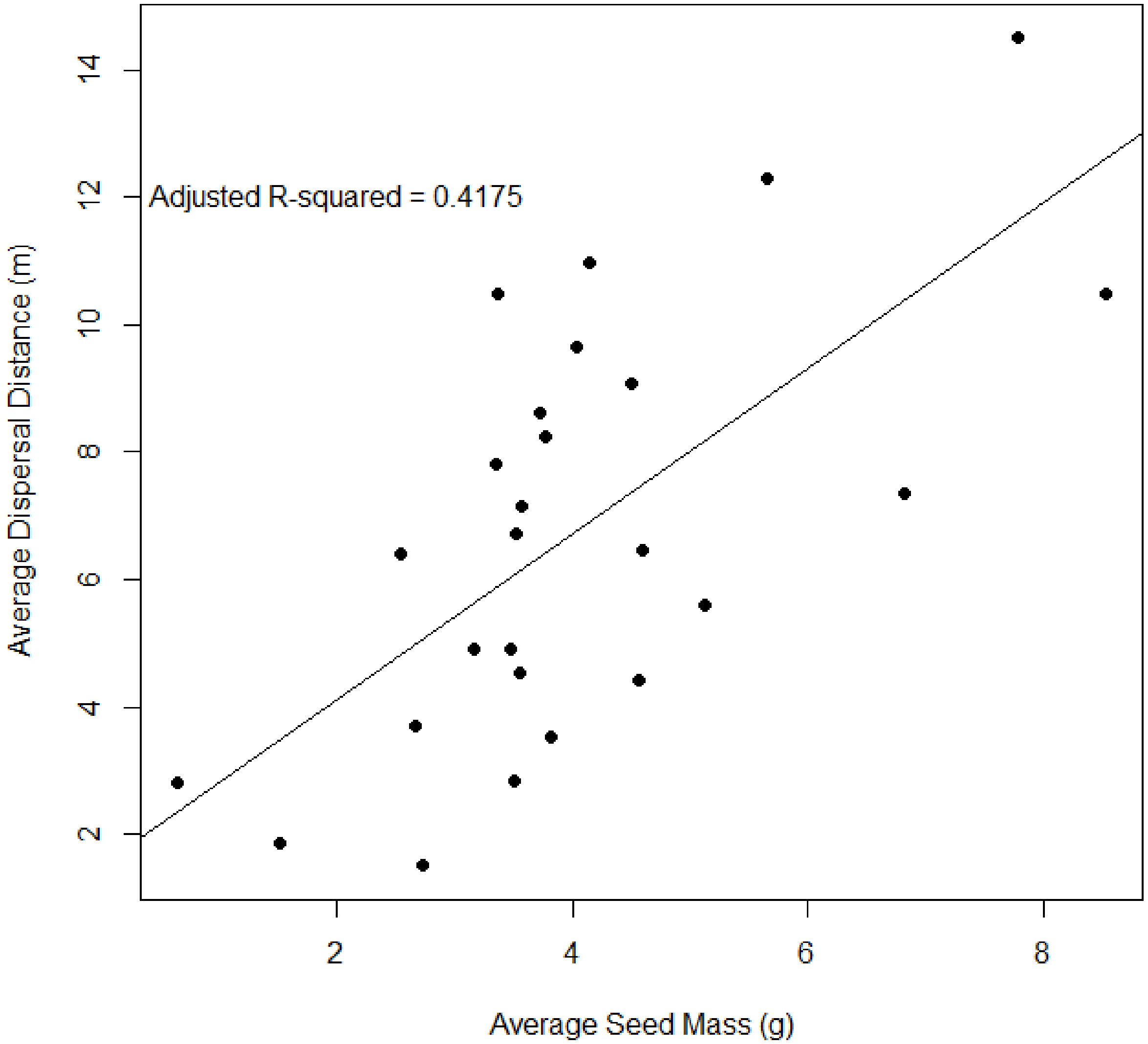


Figure 4.1 Scatterplot with simple linear regression line of average distance to caching (m) over average seed mass (g) for 25 BC3, 2 *Castanea dentata*, and 2 *C. mollissima* mother trees measured 2014-2016.

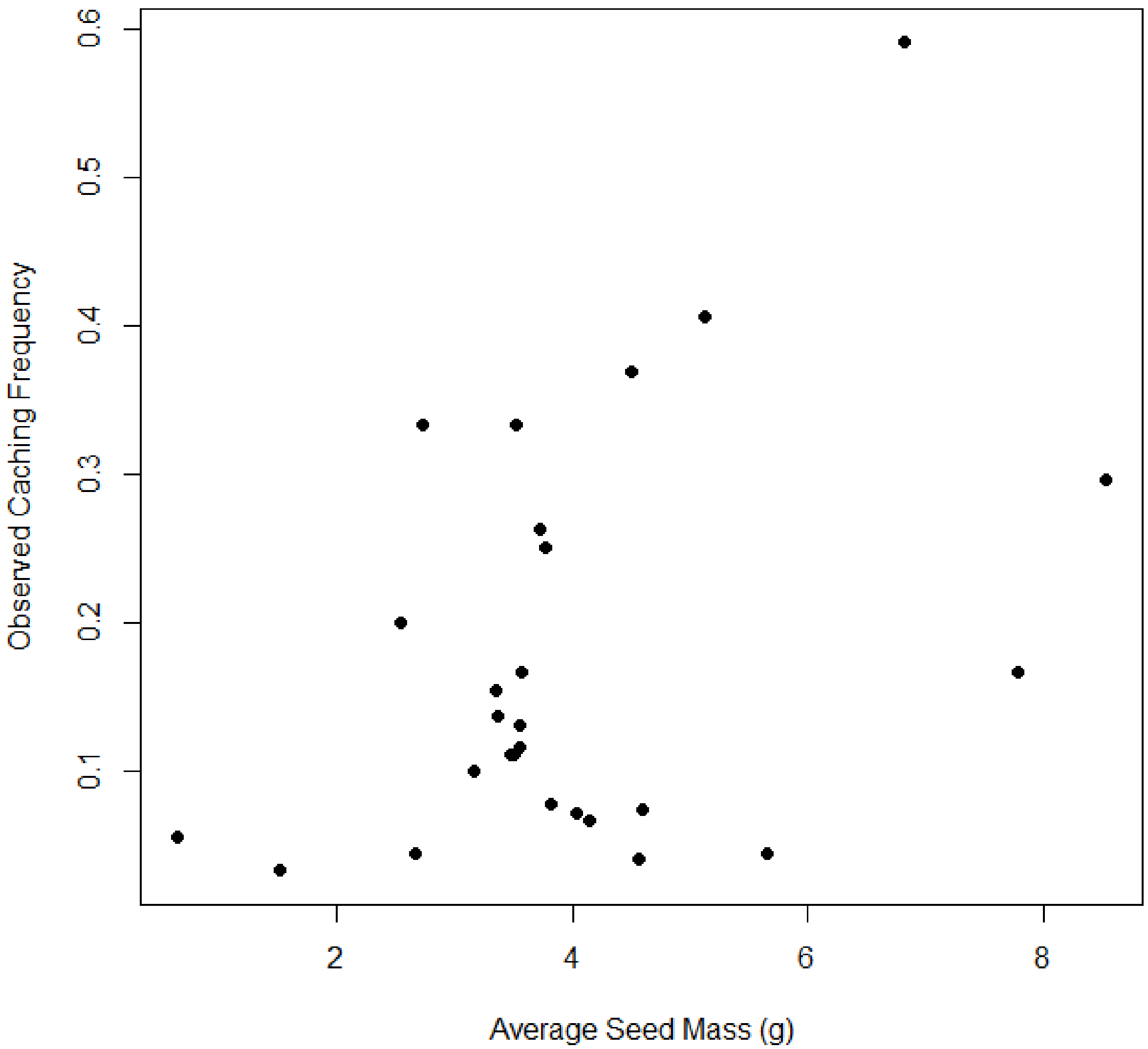


Figure 4.2 Scatterplot of percent seeds recovered in caches over average seed mass (g) for 25 BC3, 2 *Castanea dentata*, and 2 *Castanea mollissima* measured 2014-2016.

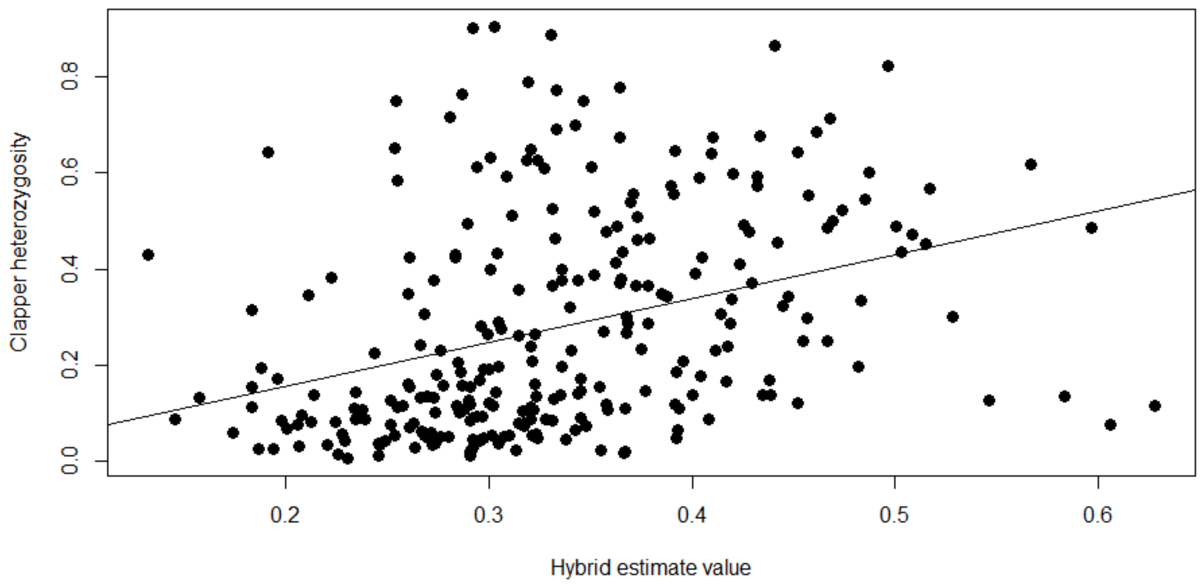


Figure 4.3 Plot of hybrid estimate value from pooled sequencing of strong-dispersal, moderate-dispersal, and weak-dispersal pools for the first 268 predicted genes on the first pseudochromosome sequence (LGA) versus heterozygosity estimates for the same genes derived from the whole-genome sequence of "Clapper."

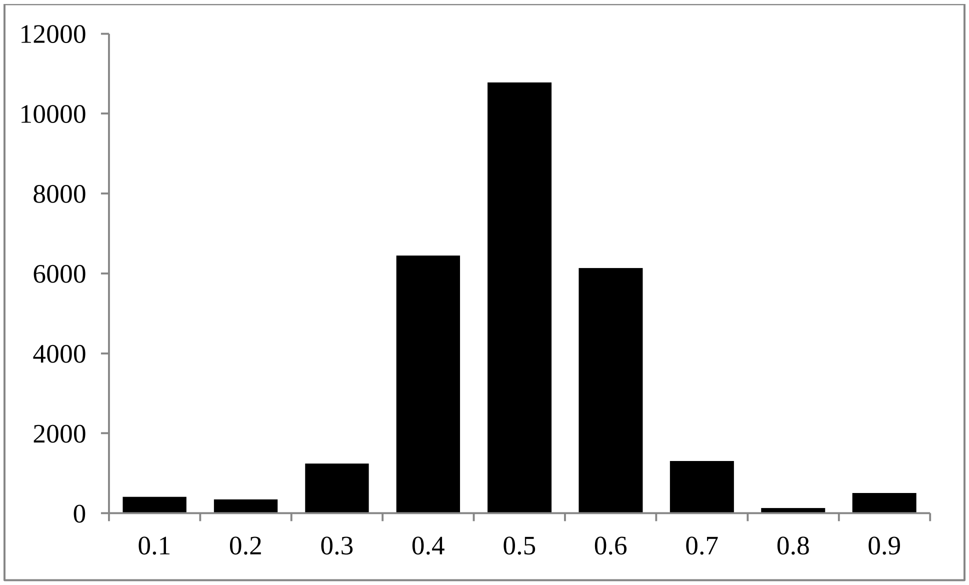


Figure 4.4. Distribution of hybridity estimator (HE) scores for pool A over all predicted genes, based on pool frequencies of *Cm* and *Cd* alleles, Pool A contained seven BC3 and one *Cm.* Y axis shows number of predicted genes in each HE score bin.

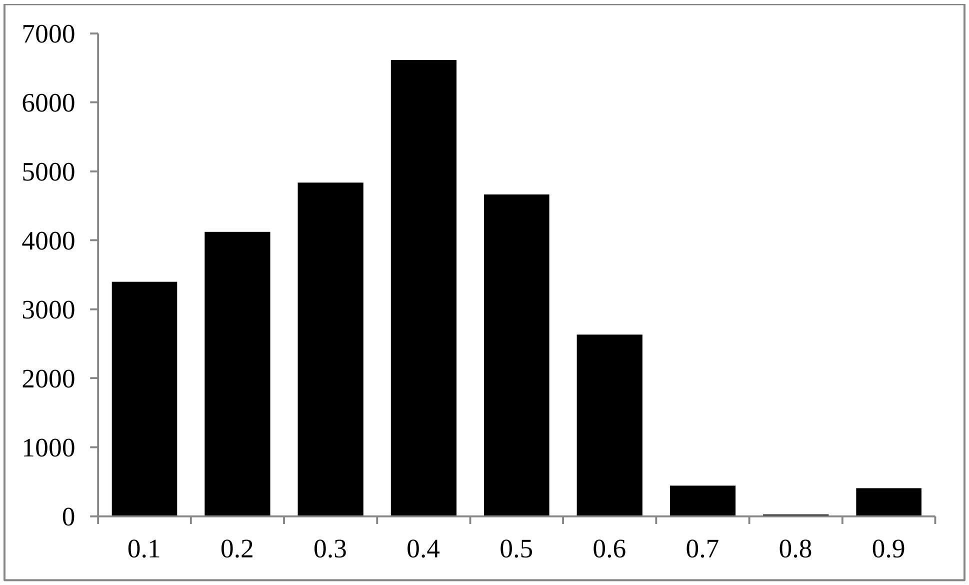


Figure 4.5 Distribution of hybridity estimator (HE) scores for pool B over all predicted genes in the genome, based on pool frequencies of *Cm* and *Cd* alleles which contained seven BC3 chestnuts. Y axis is number of genes in each HE score interval.

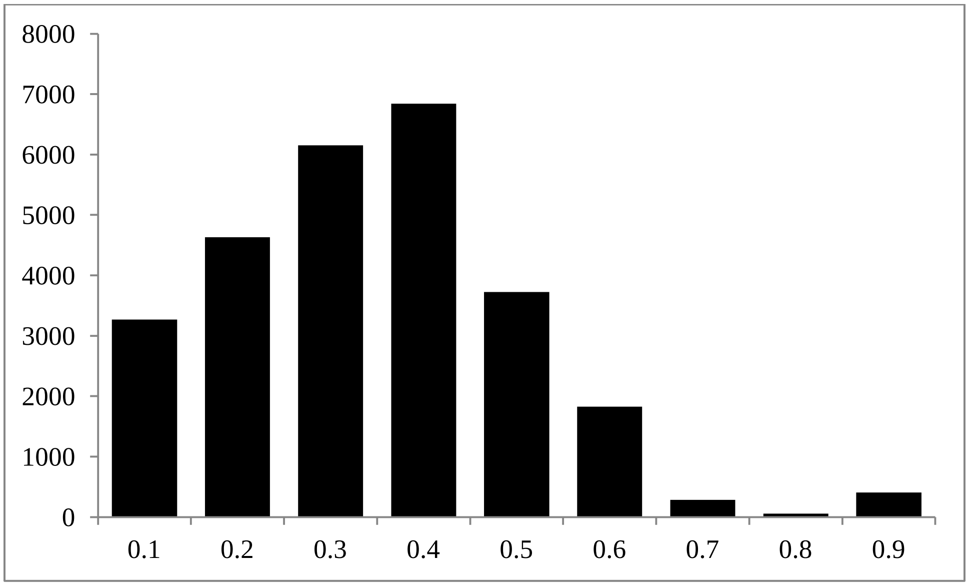


Figure 4.6. Distribution of hybridity estimator (HE) scores for pool C over all predicted genes in the genome, based on pool frequencies of *Cm* and *Cd* alleles. Pool C contained nine BC3 chestnuts and one *Cd.* Y axis is number of genes in each HE score interval.

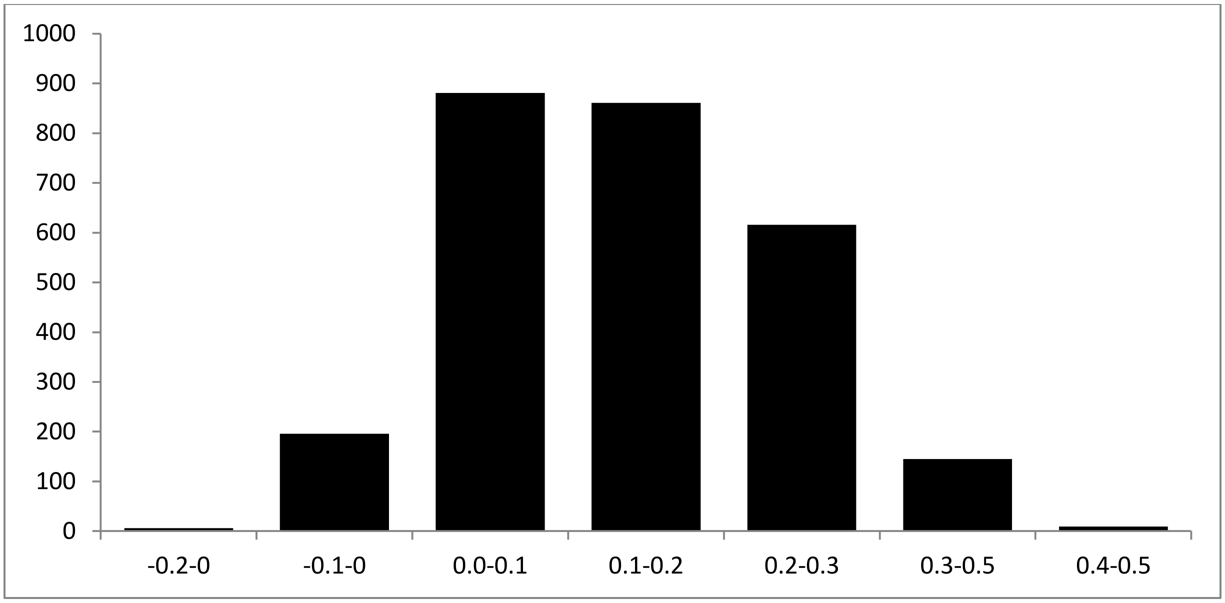


Figure 4.7 Distribution of the difference between pool A and pool B heterozygosity estimators, averaged over 10-gene bins for the entire genome. Y-axis depicts the number of 10-gene bins in a given range of the difference.

CHAPTER 5. REGIONS OF THE CHINESE CHESTNUT GENOME UNDER ARTIFICIAL SELECTION IN ORCHARD POPULATIONS AND UNDER NATURAL SELECTION IN DIFFERENT REGIONS OF CHINA

Abstract

Regions in the genome of crop plants where nucleotide diversity is reduced, relative to the rest of the crop genome and to the same region in the genomes of wild relatives, point to coding loci subject to artificial selection during the domestication process. Many of these loci, known as selective sweeps, have been identified in annual crops and perennial fruit crops, such as apple and peach, but they remain largely unknown in trees grown as nut crops. Chestnuts (*Castanea*) are major nut crops in East Asia and southern Europe, and are unique among temperate nut crops in that the harvested seeds are starchy rather than oily. Chestnut species have been cultivated for three millennia or more in China, so it is likely that artificial selection has affected the genome of orchard-grown chestnuts. The genetics of Chinese chestnut (*Castanea mollissima* Blume) domestication are also of interest to breeders of hybrid American chestnut, especially if the low-growing, branching habit of Chinese chestnut, an impediment to American chestnut restoration, is partly the result of artificial selection. We assembled genome sequences for wild and orchard-derived Chinese chestnuts and identified selective sweeps based on whole-genome SNP datasets. We present candidate gene loci for chestnut domestication and discuss the phenotypic effects of candidate loci, some of which may be useful genes for chestnut improvement in Asia and North America.

5.1 Introduction

The genomes of crop plants carry the genetic signatures of human selection, and genomic regions associated with domestication have been identified for several species. In monocotyledonous grain crops, genes considered pivotal for domestication include rice genes that affect yield by altering cell number in a part of the female flower (Shomura et al. 2008) and carbon allocation during grain filling (Wang et al. 2008), genes that control growth form (Clark et al. 2006, Tan et al. 2008, Zhou et al. 2009), seed coverings (Wang et al. 2005), starch composition of grains (Olsen et al. 2006), flowering time (Cockram et al. 2007), genes that retard shattering, the natural seed-dispersal mechanism of grasses (Doebley 2006, Li et al. 2006), and pleiotropic genes that influence several of the above traits (Simons et al. 2006). While genes related to plant architecture and grain yield were likely selected deliberately by early farmers, other loci related to domestication were probably selected unintentionally because they improved fitness of plants in cultivation regardless of their explicit appeal to humans (Zohary 2004). Annual crop plants in general display what has been called the "domestication syndrome", but which genes, and the number of genomic regions under selection, may differ based on the uses of the crop (e.g., fruits vs. seeds). Genes thought to be under selection during the domestication of legumes include seed weight, seed size, plant architecture, and flowering time in *Phaseolus vulgaris* (Schmutz et al. 2014), *Vigna angularis* (Kaga et al. 2008) and *Glycine max* (Lam et al. 2010, Li et al. 2013). Flowering time genes are also thought to have been selected by humans in the domestication of annual sunflower *Helianthus annuus,* an oilseed crop (Blackman et al. 2011). In annual dicotyledenous vegetable crops raised for their fruits (e.g. squash and tomatoes), domestication genes include several that influence fruit color (Ronen et al. 2002, Lefebvre et al. 1998), size (Frary et al. 2000), and ripening (Rao and Paran 2003, Vrebalov et al. 2002) as well as plant architecture (Mao et al. 2000, Paran and van der Knaap 2007), seed dormancy, disease resistance (Qin et al. 2014), and flavor (Guo et al. 2013, Qi et al. 2013). In the Solanaceae, which includes tomato, pepper, eggplant, and several other crops, some orthologous genes appear to have been involved in the domestication of multiple species (Doganlar et al. 2002).

Signatures of selection in the genomes of woody perennial crops may be weaker due to longer generation times and widespread self-incompatibility (Cornille et al. 2012). In the genomes of domesticated ornamental and edible peach, signatures of selection appear near genes related to flavonoid biosynthesis (flower and fruit color) and carbohydrate metabolism (fruit flavor and aroma) (Cao et al. 2014) as well as stress tolerance (Akagi et al. 2016); genes postulated to be involved in fruit development also showed signatures of selection in domesticated apple (Khan et al. 2014).

Chestnut (primarily *Castanea mollissima*) has undergone selection as a food plant in China for at least 2000 years (Jiangsu 1979, Rutter et al. 1991), a more recent event than the domestication of apples, which is estimated to have occurred about 4000 years before present (ybp) (Cornille et al. 2011) or of peach and almond, which occurred about 5000 ybp (Velasco et al. 2016). It is possible that humans began artificially selecting chestnuts earlier than 2000 ybp. Chestnuts have been found in archaeological sites dating to 6000 ybp (Jiangsu 1979) and an increase in chestnut pollen, at the expense of conifers, is noted in the archaeological record of northwest China around 4600 ybp, which coincides with the appearance of grain cultivation (Li et al. 2007). Thus, in addition to being an important food source for early Chinese civilization, chestnuts were deliberately cultivated by humans in the earliest history of China.

Today, chestnut is an economically valuable crop in east Asia, and China is the world's largest producer by far, growing large numbers of chestnuts for domestic consumption and export to Japan (Metaxas 2013). Chestnut orchards in China include both seedling trees and grafted cultivars, mostly of *C. mollissima*, with some regional use of *C. henryi*, *C. crenata*, or interspecific hybrids (Jiangsu 1979). Chestnut trees begin to bear nuts about 5 years after grafting or planting (Jiangsu 1979). Male and female catkins are produced on the same tree, but self-pollination does not normally occur (Pereira-Lorenzo et al. 2016). The timing of flower development, pollination, and fertilization of ovules is crucial for optimizing chestnut yield (Shi and Stoesser 2005). Sought-after characteristics in Chinese orchard chestnuts, for which improvement and cultivar development is ongoing, include attractive (shiny) appearance of nuts, early maturation and bearing, stable yield, high sugar content, pest and disease resistance, and adaptation to orchard environments that are hotter and drier than the mountains where most wild *C. mollissima* occur (Zhang et al. 2010). Shorter catkins are also desired (Huang et al. 2009), and large seeds(~ 20g) are sought for industrial processing into paste and flour (Xu et al. 2010). A pellicle that is easy to peel is sought after by breeders (Takada et al. 2012). Post-harvest diseases that destroy chestnuts in storage are a major concern (Ma et al. 2000).

Chestnut is somewhat unusual among orchard crops in that it is grown for starchy nuts (60-85% carbohydrates; Rutter et al. 1991) rather than oily nuts (e.g. walnut and pecan) or fleshy fruits, so it is possible that some genes under selection during chestnut domestication may be related to starch composition, as in grain crops (Olsen et al. 2006).

Plant architecture-related genes may also have been important; a small, branchy tree is more manageable in an orchard setting than a very tall one, especially in China where chestnuts are frequently picked by hand after climbing the tree (Rutter et al. 1991). Chinese chestnut in general has a shorter stature and less-pronounced apical dominance than the non-domesticated American chestnut (Clapper 1954) which is a major consideration in the backcross blight resistance breeding program being carried out by the American Chestnut Foundation and its cooperators (Burnham et al. 1986), which seeks to restore blight-resistant chestnuts with the tall stature of American chestnut. In forest settings, *C. mollissima* grow to 20-25 m in height (Fei et al. 2012), so the short stature of orchard trees may be, at least in part, an artificially selected trait. Chestnuts are highly perishable (Rutter et al. 1991) so genes related to pericarp thickness and wax coatings on the pericarp may be important if they confer improved storage qualities. Fruit quality genes, while they may not affect the flavor of the chestnut, could be under selection for human aesthetic preferences; Clapper (1954) noted variation in the color of Chinese chestnuts that was not seen in American chestnuts. Finally, although preference for large seeds varies across China (Jiangsu 1979), seed size is a likely cause for artificial selection in Chinese chestnut, especially for “processing” varieties intended for the industrial production of paste and flour (e.g. Xu et al. 2010). Finally, given the importance of the Asian chestnut gall wasp (*Dryocosmus kuriphilus*) as a pest of commercial chestnut crops worldwide (Jiangsu 1979, Rutter 1991), the many other insects that attack chestnut (Gaoping et al. 2001) and the pre- and post-harvest diseases that can affect chestnut yield, genes involved in constitutive defenses and defensive responses against insects and pathogens might also be expected to show signatures of selection in cultivated chestnuts.

In addition to differentiation between cultivated and wild Chinese chestnut, there is likely to be functional genetic differentiation among regional subpopulations of wild trees; even in orchard trees there is considerable regional variation in nut characteristics (Yang et al. 2015). Chinese chestnut occupies a larger range than any other Asian or American species of *Castanea* (Fei et al. 2012). Although this range has almost certainly been expanded by human activity, the species appears to be broadly adapted. The natural selective pressure on Chinese chestnut populations is likely to vary considerably between its temperate, high-altitude habitat in the Qin Mts (northwest China) and the subtropical provinces of Yunnan and Guizhou. Considerable rangewide genetic variation, at the whole-genome scale, has been identified in forest tree genomes, including poplar (Slavov et al. 2011) and whitebark pine (Syring et al. 2016). Genetic diversity of wild Chinese chestnut has been analyzed with varying results; either southwest (Zhang and Liu 1998) or northwest China (Shaanxi Province; Cheng et al. 2012) is likely to be the center of genetic diversity for the species. While genetic diversity is higher in wild trees, it appears that a high level of genetic diversity has been maintained in orchard (domesticated) Chinese chestnuts (Pereira-Lorenza et al. 2016), although the genetic diversity of new cultivars may be lower than traditional orchard trees (Ovesna et al. 2004).

If genetic diversity of orchard Chinese chestnuts has been lowered due to artificial selection, it should be possible to identify the genomic regions where selection has been most intense. Signatures of selection due to domestication are generally identified as regions of the genome where, for statistics related to nucleotide diversity and heterozygosity (Tajima's D, pi, FST), the differences between domesticated and wild-type lineages are greatest. Since selection theoretically leads to a loss of allelic diversity as the selected allele becomes fixed, these “selective sweeps” are regions where allelic diversity is much lower in domesticated plants than in wild plants. Typically, when whole genomes are analyzed using this type of analysis, dozens or hundreds of relatively small genomic intervals show evidence of a selective sweep. Given the large number of polymorphic sites in plant genomes, the likelihood that sweeps will be observed by chance alone (false positives) is high. However, the use of multiple statistical tests can ameliorate this problem, as can the use of non-parametric methods like permutation- or Bayesian-based tests. Genes identified in domestication regions, if subsequent investigation confirms their predicted function and phenotypic effects, could be important for further improvement of Chinese and other chestnut species for orchard production. My main questions were: 1) Is genetic diversity on the genomic scale lower in orchard-derived Chinese chestnut than it is in wild Chinese chestnut? 2) What regions of the genome show evidence of selective sweeps in the genome of domesticated Chinese chestnut, and 3) Are their regions of the genome that show different signatures of selection in northern (Shaanxi Province) and southern (Yunnan and Guizhou) gene pools of wild Chinese chestnut? To answer these questions, I utilized whole-genome resequencing with a pool-seq approach. Because I investigated genetic differentiation among different groups (pools) of trees rather than individual trees, it was feasible to estimate allele frequencies and genetic statistics (like pi and Tajima's D) from pools of samples rather than individual genome sequences. Since there were no individual phenotypes (e.g. disease resistance, seed size) available for any of the Chinese samples, there was not a need to sequence each sample individually. The advantage of this was that the sequencing cost per individual was less, so more individuals (a larger sample of the total genetic variation among wild and orchard trees) could be used to estimate population genetics statistics (Schloetterer et al. 2014, Chen et al. 2016). The drawback to this approach is that results must be interpreted with caution as false-positive rates can be high with small sample sizes (Lynch et al. 2014). Because of this uncertainty, I validated candidate loci for selection under domestication by comparing nucleotide diversity statistics and heterozygosity from individual genome sequences of 17 orchard-derived Chinese chestnuts.

5.2 Materials and Methods

5.2.1 DNA samples

Leaf samples were collected in China, rapidly dried using desiccant beads, and mailed to Purdue University for DNA isolation. Trees classified as wild were sampled from natural forests in mountainous areas where it is relatively unlikely that groves of chestnut represent escapes from cultivation (Figure 1) and ranged from Fengqing County, Yunnan, only 100 km north of the border of Myanmar, to the Heihe Forest Preserve near Xi’an in Shaanxi Province. Orchard trees were sampled from orchard settings in northeast China where most commercial growing takes place (Table 1). The United States sample of orchard-derived Chinese chestnut was obtained from Greg Miller (Empire Chestnut Company, Carrollton, OH), but the original source of the material was Beijing. DNA for these samples was isolated from dormant twigs. For leaf and twig samples, tissue (about 16 cm^2 of leaf or a 6 cm section of twig with buds) was ground to a fine powder in liquid nitrogen using a mortar and pestle, then added to a tube of heated (55 C) CTAB extraction buffer and incubated for 4-6 hours. Following incubation, DNA isolation was performed in 15 mL conical tubes using a phenol-chloroform extraction protocol, and DNA was precipitated in .2 M sodium chloride and isopropanol. After pelleting and resuspension of DNA in TE buffer, samples were cleaned using OneStep PCR Inhibitor Removal kits (Zymo Research, Irvine, CA, USA). Samples were quantified and quality assessed using a NanoDrop 8000 (Thermo-Fisher Scientific, Waltham, MA, USA) prior to pooling. Samples were pooled by source location at equimolar concentrations at a final volume of 200 uL and submitted for sequencing.

5.2.2 DNA Sequencing and Assembly

Sequencing of 100 bp paired-end reads was carried out with an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA) at the Purdue Genomics Core Facility. In order to obtain ~1x genome coverage per individual sample in a pool, six pooled samples were sequenced per lane. Low-quality reads were filtered prior to assembly using Trimmomatic version 0.32 (Bolger et al. 2014).

Chloroplasts were assembled by assembling short reads to the complete Chinese chestnut chloroplast reference sequence (Jansen et al. 2011). The 1.0 version of the Linkage Group A (LGA) pseudochromosome assembly and beta versions of the LGB-LGL assemblies (12 total) were obtained from Dr. John Carlson of Penn State University. Short reads were assembled to reference sequences using BWA, duplicates were flagged and alignment files sorted using Picard Tools, and SNPs were called using the HaplotypeCaller tool from the Genome Analysis ToolKit (GATK), with a polyploid value equal to the number of individuals in the pool. The Samtools mpileup tool was used to generate pileup-formatted SNP files for the orchard and wild sets of sample pools.

5.2.3 Identification of regions under selection in the genome

Tajima’s D and pi were calculated from mpileup files of orchard and wild assemblies using PoPoolation 2.0 (Kofler et al. 2011) over 10 kb windows for the entire genome. The difference in Tajima’s D between orchard and wild pools was calculated and statistical significance tested using a permutation test encoded in a custom Perl script. Permutations were performed by assigning observed Tajima’s D values, within the orchard and wild pools of samples, to a random base-pair interval of the genome and recalculating the difference in Tajima's D between pools over the shuffled intervals. A p-value was assigned to each interval based on how many times a difference larger than the difference at that interval was observed in 1000 shuffled genomes. Candidate loci for selection in orchard trees were intervals where the permuted p-value was less than 0.01. A second method for identifying regions in the genome under selection identified predicted gene intervals where the percent of SNPs that had one allele fixed was higher in one sample than in the other. The frequency of the major allele at SNP loci was averaged over all SNPs in a given predicted gene, and then the average major allele frequency was calculated for 10-gene intervals across the genome. Loci potentially under selection in orchard trees were identified based on the empirical distribution of the difference in the allele-frequency statistic over all predicted genes that had alignments to the UniProt database. A predicted gene was determined as potentially under selection if the difference in average major allele frequency between wild and orchard samples was greater than three standard deviations above the mean difference for all predicted genes in the genome. This method was used to identify genes under selection in orchard vs. wild trees, and also to identify loci with varying allele frequency among regional subpopulations of wild trees: northern (Shaanxi) versus southern (Yunnan + Guizhou). To reduce the false positive rate, we only considered for further analysis intervals where multiple consecutive 10kb intervals showed significantly different ($p < 0.01$) values for Tajima's D and pi in orchard versus wild trees, and/or a p-value less than 0.001.

5.2.4 Gene prediction and filtering

De novo gene prediction was carried out using AUGUSTUS (Stanke et al. 2006) with *Arabidopsis thaliana* as the training protein set and default settings. To assign a putative function to predicted genes, the predicted gene file (.gff) was converted to fasta (.fa) format and aligned to the UniProt protein database using the blastp function of the DIAMOND sequence aligner (Buchfink et al. 2015) using default settings. The top hit was assigned as the putative function of the gene.

To provide a measure of validation to this predicted gene set, publicly available cDNA contig files for American chestnut, Chinese chestnut, European chestnut, and Japanese chestnut were downloaded from http://www.hardwoodgenomics.org/transcriptomes. These were each aligned using the blastx function of DIAMOND, using default settings, to a database created using the predicted Chinese chestnut protein set output by AUGUSTUS. Transcripts were matched to the protein that provided the top hit from the predicted protein set; a predicted protein was only counted as having transcript support if it was the best alignment for at least one cDNA contig. This was carried out using a custom Perl script. The list of alignments was also searched for cDNA contigs that were designated differentially expressed in Barakat et al. (2012). AUGUSTUS output was also converted to .bed format and used to filter .vcf files (VCFtools) for those polymorphisms that occurred in predicted gene sequences and in predicted exons. Genes predicted using MAKER for LGA were downloaded from the Hardwood Genomics website for comparison of gene prediction with AUGUSTUS.

5.2.5 Identification of chloroplast haplotypes

Chloroplast reads from whole-genome sequence data were assembled to the reference Chinese chestnut chloroplast genome using BWA and Picard Tools and SNPs were called using GATK with ploidy set equal to 10. Using a custom Perl script, the number of SNPs with all possible allele frequencies (0/10 -> 10/10) was calculated for each pool to estimate the number of chloroplast haplotypes present in each regional pool. Alternate chloroplast haplotypes were identified by peaks on a histogram of SNPs in allele frequency bins for each sample; the frequency of a haplotype was estimated by the bin where a "peak" occurred, and the haplotype identity estimated by the number of SNPs in an allele frequency bin (Figure 2). SNPs were compared with individual chloroplast sequences from Chinese chestnuts (Chapter 3) to determine whether haplotypes matched either of the two previously identified haplotypes.

5.2.6 Validation of regions under selection

Whole-genome sequences of individual chestnuts (Chapter 3) were used to provide validation of regions under selection identified using pooled sequences. Tajima's D, nucleotide diversity, heterozygosity, and pi were calculated (VCFTools) using SNPs within exons of predicted genes for 18 Chinese chestnuts of southern Chinese and Korean provenance, as well as 2 American chestnuts and 4 hybrids, which represent non-domesticated trees. A negative value of Tajima's D, and low values for pi and proportion of heterozygous loci for a given predicted gene among individual orchard-derived Chinese chestnuts was interpreted as support for that gene's selection during domestication.

5.3 Results

5.3.1 Genome sequencing and assembly

Average estimated genome coverage for the pools sequenced was close to 1x per individual tree in a pool for most of the sequenced pools (Table 1) and was greater than 7x for all but two of the pools sequenced. The number of polymorphisms with alternate allele frequencies >0.2, which are less likely to result from sequencing errors, was highest in the Shaanxi orchard sample and lowest in the Beijing-derived orchard sample from Ohio (Table 2). The genomes of most of the orchard samples had fewer polymorphisms than wild trees. This could be a result of lower genetic diversity in orchard trees. The reference genome for Chinese chestnut was sequenced from the orchard cultivar 'Vanuxem' (Carlson et al. 2017), so it is possible that fewer SNPs were observed in orchard samples because they are more similar to the reference, and contain less genetic variation, than the wild samples. However, fewer DNA reads were obtained for some of the orchard sample pools than for any of the wild pools, so it is also possible that the reduced number of SNPs identified in orchard samples is partly an artifact of SNP calling procedures, i.e., polymorphisms may exist in the genomes of the orchard pools that were not included in the final dataset because coverage was too low at that site to call a SNP. Since the same minimum coverage filter (8x) was implemented for the SNP sets from orchard and wild pools during data analysis, the lower coverage in the orchard samples should not bias the identification of regions with lower genetic diversity in orchard trees, because only SNPs with adequate coverage in both samples were included in the analysis.

5.3.2 Regions under selection

Tajima's D was used as a measure of selection pressure, and Tajima's D was, on average, lower in orchard trees (-0.64) than in wild trees (-0.50). Using the Tajima's D and pi outlier method in PoPoolation, >100 intervals were significantly different between wild and orchard trees, as determined by permutation tests with a significance cutoff of p <0.01 for a given 10,000 base-pair interval. The major allele frequency across predicted gene sequences for orchard chestnuts was slightly higher (0.693) than for wild chestnuts (0.685). Using the allele frequency method to identify regions under selection, the standard deviation of the difference in major allele frequency between orchard and wild pools was used to identify outliers (cutoff: >3 standard deviations greater than mean difference for orchard vs. wild and >2 sd for regional differences), which led to the identification of approximately 25 candidate loci for domestication and 15 for regional genetic differences (Table 4, Table 5). The identified candidate loci contained predicted flowering-time genes, genes involved in the synthesis of ethylene, genes influencing male fertility, cell wall structure, secondary metabolites, and disease resistance (Table 3, Table 4, Table 6). Using whole-genome sequences of 17 individual Chinese chestnuts and 3 individual American chestnuts, we were able to determine that the candidate loci under selection showed lower-than average heterozygosity and nucleotide diversity in Chinese chestnut and, in many cases, greater nucleotide diversity in American chestnut than Chinese chestnut.

5.3.3 Chloroplast haplotypes

There was evidence of multiple chloroplast haplotypes in all but one of the populations sampled (one of the Yunnan samples; Figure 3. The reference chloroplast haplotype was found at its highest frequency in one Yunnan sample (100%) and the Guizhou sample (~60%), and at its lowest frequencies in the Hebei and ECC orchard samples (~10%). One alternate haplotype was present in the Guizhou (~40%), Hebei (~90%), ECC (90%), Beijing (~20%) and Shaanxi-3 (~90%) pooled samples (Figure 3, Figure 4). This haplotype, which had about 260 SNP polymorphisms different from the reference, was found to be the same as the (non-reference) *C. mollissima* chloroplast of ‘Clapper’ (see Chapter 3). Other polymorphic sites did not correspond to the ‘Clapper’haplotype, so other haplotypes must have been present in some of the sampled populations. A highly divergent (1000+ SNPs different from reference) haplotype appears to be present at relatively low frequency in the Shaanxi-1, Shaanxi-4, and Yunnan-2 samples (Figure 4), and an additional haplotype with low divergence from the reference, about 75 SNPs, appears to be present in the Shaanxi-1 sample (Figure 3).

5.4 Discussion

5.4.1 Chloroplast assemblies and genetic diversity

The chloroplasts assembled from pool-seq data indicated the presence of chloroplast haplotypes in Chinese chestnuts from China that were not found in any of the sampled Chinese chestnuts from the United States in Chapter 3. The most abundant haplotype in orchard-derived Chinese chestnut from China was only found in 'Clapper' among all the Chinese chestnuts sampled from the U.S. This supports the assertion that most U.S. Chinese chestnut germplasm was derived from southern Chinese source material (Rutter et al. 1991). The Shaanxi orchard chestnut sample's chloroplast genotype profile resembled the wild Shaanxi-1 chloroplast profile more than it did the other orchard samples, which indicated that admixture between local wild populations and orchard trees is probably extensive in cultivated Chinese chestnut. Given that Chinese chestnuts are commonly cultivated in mountainous areas where wild *C. mollissima* is also most likely to occur, this makes sense. The traditional practice of growing seedling chestnuts would also favor admixture, although the presence of Shaanxi-specific chloroplast haplotypes in the Shaanxi orchard populations indicated maternal inheritance, i.e., that admixture is not the result of local pollen alone. The large number of SNPs identified in the nuclear genome of the Shaanxi orchard sample (Table 3) also indicated extensive admixture in this population. The chloroplast haplotype shared by 'Clapper' and two of the orchard pools (Hebei and ECC) was also found at high frequency in the Shaanxi-3 wild sample. Xi'an, one of the earliest civilizational centers of China, is located in Shaanxi province, so it is possible that chestnut was first brought into cultivation there and later spread to the area around Beijing. It is also possible that this haplotype is more common in wild trees north of the Qin Mountains and that (as in the Shaanxi orchard sample) orchard trees were selected from the local chestnut gene pool. The diversity of chloroplast haplotypes evident in the three wild Shaanxi samples corroborates earlier findings that the Qin Mountains (Shaanxi province) represent a center of genetic diversity for *C. mollissima* (Cheng et al. 2012). More sampling is needed to parse out how many haplotypes are present in the Shaanxi and Yunnan chestnut populations, where the best evidence for diversity was observed.

Previous studies of genetic diversity in wild and orchard Chinese chestnuts found higher genetic diversity in wild trees, but relatively high genetic diversity maintained in orchard trees (Pereiera-Lorenzo et al. 2016). It appears to be the case that, like other perennial woody food plants (Cornille et al. 2011) the overall reduction in genetic diversity due to domestication has been limited. Despite this, the number of relatively small (50-100 kb) regions in the genome where orchard trees had depressed genetic diversity relative to wild trees was fairly large, and about 10x larger than the number of regions where orchard trees had elevated genetic diversity relative to wild trees. By looking at nucleotide diversity statistics for some candidate genes supported by high-quality data (>10x coverage) from individually sequenced genomes of 17 orchard-grown Chinese chestnuts, we were able to identify several loci that showed strong evidence of lowered genetic diversity (data from Chapter 3). These loci showed lower genetic diversity in orchard-derived Chinese chestnut versus the non-domesticated American chestnut, as well as very strong differentiation (measured by F_{ST}) between the two species. These loci (in bold text in Table 6) we consider the least likely to represent statistical artifacts and false-positives, and most likely to be candidates for chestnut domestication. For many of the loci with putative selective sweeps, gene annotations were similar to genes in putative domestication selective sweeps from other crop species, and corresponded with expectations about the traits under selection in Chinese chestnut: nut size, flowering time, tree architecture, defense, traits related to nut ripening and dropping, and secondary metabolites that may impact pest resistance, storage quality, aesthetic appeal, and flavor of chestnuts.

5.4.2 Inferred roles of regions under selection in chestnut domestication

Some genes in putative sweep regions appeared to be directly involved in the processing of secondary compounds that could affect the flavor of chestnuts. One, on LGD, was similar to a flavonol synthase gene from *Citrus unshui* (Lukacin et al. 2003) that is involved in the synthesis of several flavonoid compounds. Flavonoids are secondary compounds that have an influence on the bitter flavors present in citrus

(Frydman et al. 2013) and tea (Xia et al. 2017), among other food plants. The linkage group A pseudochromosome reference sequence contained a predicted gene in a putative sweep region that wassimilar to the dioxygenase AOP1 of *Arabidopsis*, which is probably involved in glucosinalate synthesis (Kliebenstein et al. 2001). Another potential selective sweep on LGA contained a predicted gene similar to anthocyanidin 3-O-glucosyltransferase 2 of wine grapes (*Vitis vinifera*), which is responsible for the synthesis of red wine pigments (Ford et al. 1998) It is plausible that secondary metabolites in general would be selected against during domestication because they affect the flavor of chestnuts, but a pigment gene might be selected *for* if it confers a pleasing red color in the chestnut shell or leaves; there are Chinese chestnut cultivars with enhanced red coloration in their leaves and twigs (Junhao et al. 2000). Secondary metabolites might also be selected for if they provide some benefit in terms of insect or disease resistance. Chestnuts with excess red pigmentation in leaves, twigs, and nuts may be less vulnerable to post-harvest fungal infection (G. Miller, pers. comm.).

Flowering time genes are among the most frequently identified in selective sweeps related to domestication (e.g. Kaga et al. 2008, Schmutz et al. 2014). A crop plant, whether it is a grass or an orchard tree, must flower so that pollen is available to fertilize female flowers and maximize yield. Predicted genes similar to known flowering-time regulatory genes were found at several putative selective sweep loci. One, on LGA, was a homolog of FLOWERING LOCUS C (FLC), a MADS-box protein that functions as major floral development repressor (Choi et al. 2009); another on LGI encoded a protein similar to FTIP1 of *Arabidopsis*, which exports the essential flowering control protein FLOWERING TIME (FT) into phloem sieve elements (Liu et al. 2012). The FLOWERING LOCUS C homolog showed a particularly strong signature of selection in the 17 whole-genome sequences we obtained from orchard-derived Chinese chestnuts (Table 6). Another sweep that may involve a fertility-related gene was identified on LGD; it contained a predicted gene similar to an egg-cell-secreted protein from *Arabidopsis* that governs gamete interactions (Sprunck et al. 2012) and is also present in early embryos.

Several other selective sweeps appeared to include genes associated with flower development, in particular, with the development of male flowers. One on LGE was similar to the POLLENLESS gene of *Arabidopsis*. POLLENLESS has a crucial role in male fertility in *Arabidopsis* and mutants are typically male-sterile (Glover et al. 1998). This gene could have been under selection during chestnut domestication because trees that produce less pollen tend to produce more seeds; a short-catkin mutation of Chinese chestnut has previously been identified (Feng et al. 2011), and improvement of this trait could greatly increase yield of chestnuts. Some *Castanea sativa* cultivars with exceptionally large nuts ("marron" types) actually produce astaminate catkins that are sterile (Pereira-Lorenzo et al. 2006) and there is considerable variation in the size of male catkins among cultivars (Pereira-Lorenzo et al. 2016). Remarkably, another sweep (on LGK) contained a predicted gene homologous to the *Arabidopsis* AGAMOUS gene, a probable transcription factor that controls organ identity in developing flowers (Drews et al. 1991). Disruptions of AGAMOUS lead to loss-of-function in both male and female flowers (Yanofsky et al. 1990), so the chestnut gene apparently under selection on LGK may not be involved in male sterility. It could also regulate carpel development and therefore the ultimate size and shape of the seeds. A sweep region on LGC contains a predicted gene homologous to SUVH4 of *Arabidopsis,* which is also known as KRYPTONITE because of its molecular role in suppressing the SUPERMAN transcription regulator (Jackson et al. 2002). SUVH4 silences its targets by methylating DNA. SUPERMAN is involved in female floral development (Sakai et al. 1995), so this locus could plausibly be selected for enhanced formation of female flowers versus male flowers.

A number of the predicted genes in the regions with signatures of selection in orchard trees were similar to genes in model plants that are involved in the regulation of plant development. In chestnut, some of these genes might be involved in the distinctive low-branching phenotype of orchard-derived Chinese chestnut These include a predicted gene similar to phytosulfokine receptor PSK6 of *Arabidopsis,* which affect the longevity of cells and their potential for growth (Matsubayashi et al. 2006). One intriguing gene is similar to a shoot gravitropism regulator (SGR5 or IDD15) of *Arabidopsis,* a transcription factor that regulates branch orientation (Cui et al. 2013) and starch levels (Tanimoto et al. 2008). A putative sweep region on the LGB pseudochromosome sequence contained a gene similar to RICESLEEPER1 from rice, a major transcriptional regulator widely distributed in angiosperms. RICESLEEPER 1 mutants show reduced size and seed production (Knip et al. 2012) so this gene could potentially influence important multiple important traits for the domestication of chestnut. A putative sweep locus on LGC contains a predicted gene whose protein product is similar to a cell-number regulation enzyme in maize that affects plant organ size and is homologous to a major fruit weight QTL gene in tomato (Guo et al. 2010). If this gene has the same effect of modifying fruit size in chestnut, it would likely have been an important factor in chestnut domestication, especially in regions where large seed size is favored. Given that starch is the major nutritional constituent of chestnuts, genes that affect starch synthesis may have been important during domestication. One locus on LGB contained a predicted gene similar to sucrose synthase 2 of *Arabidopsis*, which is involved in furnishing carbon for starch synthesis in developing seeds (Angeles-Nunez and Tiessen 2010).

Many of the intervals with putative involvement in domestication contained genes associated with cell wall development. Modification of cell walls is a major part of fruit ripening, which is why polygalacturonases, cellulases, and other cell-wall enzymes have been discovered in selective sweeps in the genomes of domesticated tomato and pepper (Paran and van der Knapp 2007). One predicted gene in a selective sweep on LGF was similar to *Arabidopsis* RABA4B, a Golgi-network trafficking regulatory protein that may involved in the secretion of cell wall components (Preuss et al. 2004). Sweeps on LGA and LGE contain predicted polygalacturonases, one of which is expressed in flowers, but is probably not involved in the final process of cell wall modification by which seed pods split open (Ogawa et al. 2009) and a polygalacturonase similar to ADPG2 in *Arabidopsis*, which is involved in pod shattering (Gonzalez-Carranza et al. 2007, Ogawa et al. 2009). One putative sweep locus, on LGD, contained a predicted gene similar to a WAT1-related protein of *Arabidopsis*, which belongs to the UmamiT amino acid transporter class; mutants of similar proteins in *Arabidopsis* show reduced seed size (Müller et al. 2015), so it may have some role in provisioning developing seeds with amino acids, or a role in cell wall formation. Thicker cell walls could improve the storage properties of chestnuts by increasing pericarp thickness. A predicted gene similar to glycerophosphoryl diester phosphodiesterase GDPDL3 of *Arabidopsis*, which is essential for the construction of the primary cell wall (Hayashi et al. 2008) was located in a selective sweep region on LGA.

Enzymes that modify and transport lipids may have been important in chestnut domestication not because chestnut is grown for lipid content, but rather because cuticular waxes likely influence the storage quality and appearance of chestnuts, as well as resistance to pests and pathogens. One putative sweep on LGB contained a fatty acid beta-oxidation protein similar to a probable enoyl-CoA hydratase from the *Arabidopsis* peroxisome (Reumann et al. 2007). Another predicted gene, in a sweep on LGG, encodes a protein similar to isocitrate lyase of cotton, a gene which in *Arabidopsis* is involved in mobilizing storage lipids during germination and growth of seedlings (Eastmond et al. 2000). At a different locus on LGG, a predicted protein in a putative sweep is similar to a peroxisome biogenesis protein of *Arabidopsis* that may be involved in fatty acid beta-oxidation (Nito et al. 2007). Putative selective sweeps on LGC and LGA are home to predicted genes that are similar to an acyl carrier protein that is probably involved in fatty acid biosynthesis and an alkane hydrolase (MAH1) that is involved in forming cuticular waxes in *Arabidopsis* (Greer et al. 2007). One predicted gene in a sweep interval on LGE is similar to a fatty acid amide hydrolase gene in *Arabidopsis* that is involved in breaking down signal-transducing fatty acids, and appears to have an effect on both growth and disease-response phenotypes (Kim et al. 2009). A selective sweep region on LGD contains a predicted gene that is similar to aromatic-L-amino-acid decarboxylase of *Homo sapiens*, which is an enzyme that synthesizes dopamine, serotonin, and tryptamine from their respective precursors (Giardina et al. 2011). The predicted gene here could be involved in the synthesis of other secondary metabolites, but serotonin and tryptamine are found in plants (Badria 2002) as is dopamine, which may be an allelochemical in some legumes (Guidotti et al. 2013). Dopamine cannot be absorbed from food, but its precursor L-Dopa can, and is found in some plant foods (Ramya and Thaakur 2007). It is more likely that this predicted gene was under selection during domestication due to some role in plant signalling or defense, but the potential for a direct neurological effect of chestnut consumption by humans is intriguing.

Management of disease and environmental stress was the inferred role of several predicted genes within the putative domestication intervals. Several appear to be involved in responses to pathogens. One gene in a sweep on LGK was similar to Xa21 of rice, a receptor kinase involved in initiating a resistance reaction to the bacterium *Xanthomonas oryzae* (Song et al. 1995). Another, on LGJ, encoded a predicted protein similar to the universal stress protein PHOS34 of *Arabidopsis*, which interacts with the mitogen-activated protein kinases (MAPKs) that activate resistance responses following the detection of oomycete zoospores, flagellin, and other elicitors (Merkouropoulos et al. 2008). Another potential sweep on LGC contains a gene similar to a general defense gene of tobacco, PDR, that is involved in elicitor response (Sasabe et al. 2002). One potentially relevant gene in a region with a signature of selection on LGJ in orchard trees was an ethylene-responsive transcription factor similar to *Arabidopsis* ERF3, which is involved in the regulation of stress responses and pathogenesis. One gene on LGA is similar to a translocator protein homolog in *Arabidopsis*, which is a membrane-bound protein expressed during osmotic stress (Guillaumot et al. 2009). One of the predicted peroxidases, in a sweep region on LGI, is similar to PER24 of *Arabidopsis*, which is up-regulated in response to cold (Fowler and Thomashow 2002). Peroxidases in general are frequently involved in stress responses (Valerio et al. 2004) and other predicted peroxidases were located in sweep regions on LGC and LGL. Pollen allergen-like proteins were predicted for two selective sweeps on LGJ and LGD; these are most likely also pathogenesis-related (Chen et al. 2006): a different cluster of predicted major pollen allergen genes (PruAr1-like) on LGJ may be associated with variation in chestnut blight resistance (Chapter 3).   A predicted gene similar to an *Arabidopsis* L10-interacting MYB domain protein was found in a sweep in LGH, similar to a gene in *Arabidopsis* that reduces the severity of virus infection by suppressing translation (Zorzatto et al. 2015). A locus putatively under selection on LGB contained several predicted sieve element occlusion proteins similar to SEOB of *Arabidopsis*. These proteins have been proposed to be involved in resistance to insects, but experimental evidence of this is lacking (Knoblauch et al. 2014).

Insect pests can be a major factor affecting yield of orchard-grown chestnuts; chestnut gall wasp is probably the most destructive, but aphids can also damage shoots, weevils (*Curculio sayi*) affect the quality of harvested nuts, and ambrosia beetles (*Xylosandrus* spp) damage stems and transmit fungal pathogens (Youngsteadt and

Gurney 2013). Several loci appeared to be involved in the synthesis of alkaloids, such as predicted genes similar to a reticuline oxidase gene from poppy (*Papaver somniferum*) (Facchini et al. 1996) on LGC and geraniol 10-hydroxylase (LGB), a gene that synthesizes precursors for terpenoids (Collu et al. 2001). Another gene (LGE) was similar to a tropinone reductase-like gene from coca (*Erythroxylum coca*) that is similar to tropane alkaloid genes from the nightshade family, but not actually involved in the synthesis of tropane alkaloids in coca (Jirchitzka et al. 2012). Alkaloids are seemingly not as abundant in Fagaceae as in most other plant groups (Li and Willaman 1968) but are known to occur in chestnut (Cho et al. 2015), so it is likely that the predicted gene here is involved in the synthesis of alkaloid compounds other than tropane alkaloids. The biological function of alkaloids (along with anthocyanins and tannins) is largely to deter herbivores and insects (Levin 1976), so it is possible that this gene was under selection due to pressure from insect pests. These genes could also potentially be involved in the synthesis of volatile compounds; Chinese chestnuts that produce fewer volatiles can be more resistant to chestnut gall wasp because they provide a weaker lure to the adult insect (Huang et al. 1990). Another gene potentially involved in insect defense is a predicted protein similar to the TIFY 6B transcriptional regulator of *Arabidopsis*, which modulates the response to wounding and herbivory by regulating the jasmonate signalling pathway (Chung et al. 2008).

Tolerance of abiotic stress was the inferred function for several genes in putative sweep regions, including several predicted genes that had homologs involved in response to osmotic stress induced by drought or salinity. One was similar to a desiccation related protein from blue gem (*Craterostigma plantagineum*) that may serve as an osmoprotecant in embryos of other species' seeds- it is similar to LEA proteins from cotton (Piatkowski et al. 1990). Another predicted gene potentially selected due to its role in stress tolerance is a predicted gene on LGH that is similar to BAG4 in *Arabidopsis,* which encodes a chaperone protein that has pleiotropic effects on stress responses as well as plant and inflorescence architecture (Doukhanina et al. 2006). Two separate putative sweeps on LGA and LGG included predicted genes with protein products similar to late-embryogenesis-abundant (LEA) proteins from orange (*Citrus aurantium var. chinensis*) and cotton (*Gossypium hirsutum*). LEA genes encode hydrophilic proteins are believed to have a role in protecting conferring desiccation tolerance to seeds and vegetative tissues (Battaglia et al. 2008). The citrus homolog of the predicted gene in the LGA selective sweep has been shown to be up-regulated during osmotic stress (Naot et al. 1995); the predicted genes I identified could have a role in the desiccation tolerance of seed and could have been selected during domestication because they influence the storage properties of chestnuts. Alternatively, they could be related to general plant stress tolerance. Two separate sweep regions contained predicted genes similar to homeobox-leucine zipper transcription factor proteins of *Arabidopsis*, ATHB-6 and ATHB-14 (also known as PHB). The former is involved in the response to water deficit and the negative regulation of ABA, and is expressed in carpels (Soederman et al. 1999) while the latter is involved in ovule development (Sieber et al. 2004). Two individual selective sweeps on different linkage groups (LGA,LGC) contained predicted genes that were similar to 1-aminocyclopropane-1-carboxylate oxidase genes from *Arabidopsis* and a third (LGL) contained one that was similar to 1-aminocyclopropane-1-carboxylate synthase. The products of these genes together regulate the production and degradation of the plant hormone ethylene (Yamagami et al. 2003, Qin et al. 2007). The 1a1c-oxidase genes in *Arabidopsis* promote stem elongation (Qin et al. 2007) so it is possible that these genes were selected because they influence height growth and form of Chinese chestnut. Given ethylene’s importance in fruit ripening, it is also possible that they play a role in the maturation of nuts.

Several classes of transcription factors have been implicated in plant domestication due to their influence on flower, fruit and seed development: the bHLH and MYB-family transcription factors in particular have been identified in domestication sweep regions of the genomes of peach (Cao et al. 2014) and apple (Khan et al. 2014), as well as other plants (e.g. Schmutz et al. 2014). Several MYB- and bHLH-type transcription factors were found in regions that showed evidence of strong selection in the genomes of orchard chestnuts. One basic helix-loop-helix (bHLH) – type transcription factor had a homolog in *Arabidopsis*, BH147, that is involved in brassinosteroid signallining (Wang et al. 2007) and another has a homolog to the *Arabidopsis* bHLH78 transcription factor, which promotes the expression of the Flowering Time gene and therefore is involved in the initiation of flowering (Liu et al.

2013). One MYB-type transcription factor identified in regions with signatures of selection is similar to MYB108 of *Arabidopsis*, which has a role in formation of male floral parts (Mandaokar and Brown 2009) as well as stress responses (Mengiste et al. 2003). Another was similar to *Arabidopsis* RAX2, which is involved in controlling axillary meristem development and inflorescence structure. This transcription factor could have been under selection during domestication because it controlled the branchiness of chestnut trees, or because it influenced the structure of the chestnut female inflorescence in some way that improved yield, perhaps by altering the number of nuts per inflorescence. Another transcription factor is similar to a member of the *Arabidopsis* GRAS gene family, a scarecrow-like protein that is expressed in sepals, stamens, pistil, and leaves (Lee et al. 2008). The transcription factor on LGJ that is similar to VRN1 of *Arabidopsis* could be important for domestication of chestnut if its function is the same as its *Arabidopsis* homolog, which delays flowering by binding other transcription factors that initiate the floral development program (Levy et al. 2002).

In some cases, loci were identified because one regional gene pool of wild trees had significantly higher major-allele frequencies than the other, although the number of these loci was small. Most such loci were closer to fixation in the southern samples of wild trees (Yunnan and Guizhou) than in the northern sample, with the exception of one interval on LGE that contained a predicted gene similar to cinnamoyl alcohol dehydrogenase from *Eucalyptus botryoides*, and another on LGH that was similar to a senescence-associated protein from *Arabidopsis*. The locus on LGE is intriguing because it may correspond to a QTL for resistance to *Phytophthora cinammomi* resistance in hybrids of Chinese and American chestnut (Olukolu et al. 2012). It is possible that more alleles for this gene are present in southern Chinese populations of chestnut to combat variable races of *P. cinnammomi,* which is more of a problem for chestnut in warm climates. Several other genes in regions with differentiated allele frequencies among regional subpopulations included several lignin-synthesis genes, and a DRE1B-type gene, all of which are probably involved in cold-tolerance. Interestingly, one predicted gene that had decreased allele frequency in southern China was similar to a transcription factor in *Arabidopsis* that controls trichome density (Schnellmann et al. 2002). Increased trichome density could be favorable in warmer climates where water loss is more severe during hot weather.

5.5 Conclusions

While it is interesting to understand the genetic basis of domestication in chestnut, the genomic loci identified here could have a practical basis for tree improvement programs. For breeders who are interested in improving Chinese chestnut for increased nut production or nut size, genes that were selected during domestication to promote heavier fruiting, such as the male-sterility genes identified here, could be a pathway to trees with shorter catkins and more female flowers. Many of the genes potentially involved in cuticular wax synthesis, stress tolerance, and synthesis of secondary compounds could be used for improving storage quality and pest resistance of chestnuts. For breeders who are interested in transferring disease resistance from Chinese chestnut into other species, genes potentially involved in orchard-type crown architecture might be desirable or undesirable, depending on the phenotypic goals of the program. Conversely, some of the genes identified in these sweep regions may be desirable for improving the resistance of other chestnut species to pests like Asian gall wasp and *Phytophthora* root rot. More research is needed to determine the actual phenotypic effects of the gene loci identified here, but our results provide a glimpse of selective pressure on the chestnut genome during the tree’s transition to a domesticated existence, and a rough sketch of a map for future genomics-assisted chestnut improvement.

5.6 Literature cited

Akagi T, Hanada T, Yaegaki H, Gradziel TM, Tao R (2016) Genome-wide view of genetic diversity reveals paths of selection and cultivar differentiation in peach domestication. DNA Res 23(3):271-282.

Angeles-Nunez JG, Tiessen A (2010) *Arabidopsis* sucrose sythase 2 and 3 modulate metabolic homeostasis and direct carbon towards starch synthesis in developing seeds. Planta 232:701-718.

Badria FA (2002) Melatonin, serotonin, and tryptamine in some Egyptian food and medicinal plants. J Med Food 5(3):153-7.

Battaglia M, Olvera-Carrillo Y, Garciarubio A, Campos F, Covarrubias AA (2008) The enigmatic LEA proteins and other hydrophobins. Plant Physiol 148(1):6-24.

Blackman BK, Rasmussen DA, Strasburg JL, Raduski AR, Brke JM, Knapp SJ, Michaels SD, Rieseberg LH (2011) Contributions of flowering time genes to sunflower domestication and improvement. Genetics 187:271-287.

Buchfink B, Xie C, Huson D (2015) Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59-60.

Cao K, Zheng Z, Wang L et al. (2014) Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. Genome Biology 15:415.

Chen J, Kaellman T, Ma X-F, Zaina G, Morgante M, Lascoux M (2016). Identifying genetic signatures of natural selection using pooled population sequencing in *Picea abies.* G3: Genes, Genomes, Genetics https://doi.org/10.1534/g3.116.028753

Chen M, Xu J, DEvis D, Shi J, Ren K, Searle I, Zhang D (2016) Origin and functional prediction of pollen allergens in plants. Plant Physiol 172(1):341-357.

Cheng L-L, Feng H-D, Rao Q, Wu W, Zhou M, Hu G-L, Huang W-G (2012) Diversity of wild Chinese chestnut chloroplast DNA SSRs in Shiyan. J Fruit Sci 3:382-386.

Cho J-Y, Bae S-H, Kim H-K (2015) New quinolinone alkaloids from chestnut (*Castanea crenata* Sieb) honey. J Agric Food Chme 63(13):3587-3592.

Choi J, Hyun Y, Kang MJ, In Yun H, Yun JY, Lister C, Dean C, Amasino RM, Noh B, Noh YS, Choi Y (2009) Resetting and regulation of Flowering Locus C expression during *Arabidopsis* reproductive development. Plant J 57(5):918-931.

Chung HS, Koo AJ, Gao X, Jayanty S, Thines B, Jones AD, Howe GA (2008) Regulation and function of *Arabidopsis* JASMONATE ZIM-domain genes in response to wounding and herbivory. Plant Physiol 146:952-964.

Clark RM, Nussbaum-Wagler T, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescence architecture. Nature Genetics 38:594-597.

Cockram J, Jones H, Leigh FJ, O'Sullivan D, Powell W, Laurie DA, Greenland AJ (2007) Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. J Exp Bot 58(6):1231-1244. 8(5):e1002703.

Collu G, Unver N, Peltenburg-Looman AM, van der Heijden R, Verpoorte R, Memelink J (2001) Geraniol 10-hydroxylase, a cytochrome P450 enyzme involved in terpenoid indole alkaloid biosynthesis. FEBS Lett 508(2):215-20.

Cornille A, Gladieux P, Smulders MJM et al. (2012) New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. PLoS Genetics

Cui D, Zhao H, Jing Y, Fan M, Liu J, Xin W, Hu Y (2013) The *Arabidopsis* IDD14, IDD15, and IDD16 cooperatively regulate lateral organ morphogenesis and gravitropism by promoting auxin biosynthesis and transport. PLoS Genet 9:E1003759-E100379.

Doebley J (2006) Unfallen grains: how ancient farmers turned weeds into crops. Science 312(5778):1318-1319.

Doganlar S, Frary A, Daunay M-C, Lester RN, Tanksley SD (2002) Conservation of gene function in the Solanaceae as revealed by comparative mapping of domestication traits in eggplant. Genetics 161(4):1713-1726.

Doukhanina EV, Chen S, van der Zalm E, Godzik A, Reed J, Dickman MB (2006) Identification and functional characterization of the BAG protein family in *Arabidopsis thaliana*. J Biol Chem 281:18793-18801.

Drews GN, Bowman JL, Meyerowitz EM (1991) Negative regulation of the *Arabidopsis* homeotic gene AGAMOUS by the APETALA2 product. Cell 65:991-1002.

Eastmond PJ, Germain V, Lange PR, Bryce JH, Smith SM, Graham IA (2000) Postgerminative growth and lipid catabolism in oilseeds lacking the glyoxylate cycle. Proc Nat Acad Sci USA 97:5669-5674.

Facchini PJ, Penzes C, Johnson AG, Bull D (1996) Molecular characterization of berberine bridge enzyme genes from opium poppy. Plant Physiol. 112:1669-1677.

Fei S, Liang L, Paillet FL, Steiner KC, Fang J, Shen Z, Wang Z, Hebard FV (2012) Modeling chestnut biogeography for American chestnut restoration. Diversity Distrib. 18:754-768.

Feng Y-Q, Shen Y-Y, Qin L, Cao Q-Q, Han Z-H (2011) *Short catkin1*, a novel mutant of *Castanea mollissima*, is associated with programmed cell death during chestnut staminate flower differentation. Scientia Horticulturae 130(2):431-435.

Ford CM, Boss PK, Hoj PB (1998) Cloning and characterization of Vitis vinifera UDP-glucose:flavonoid 3-O-glucosyltransferase, a homologue of the enzyme encoded by the maize Bronze-1 locus that may primarily serve to glucosylate anthocyanidins in vivo. J Biol Chem 273:9224-9233.

Fowler S, Thomashow MF (2002) *Arabidopsis* transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway (2002) Plant Cell 14:1675-1690.

Frydman A, Lieberman R, Huhman DV, Carmeli-Weissberg M, Sapir-Mir M, Ophir R, Sumner LW, Eyal Y (2013) The molecular and enzymatic basis of bitter/non-bitter flavor of citrus fruit: evolution of branch-forming rhamnosyl-transferases under domestication. Plant J 73:166-178.

Frary A, Nesbitt TC, Grandillo S, van der Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD (2000) *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85-88.

Fujimoto SY, Ohta M, Usui A, Shinshi H, Ohme-Takagi M (2000) *Arabidopsis* ethylene responsive element binding factors act as transcriptional activators or repressors of GCC box mediated gene expression. Plant Cell 12:393-404.

Gaoping W, Qing Y, Kai Z, Ciesla WM (2001) Factors affecting production of Chinese chestnut in Xinxian County, Henan Provinve, China. Forestry Chronicle 77(5):839.

Giardina G, Montioli R, Gianni S, Cellini B, Paiardini A, Voltattorni CB, Cutruzzola F (2011) Open conformation of human DOPA decarboxylase reveals the mechanism of PLP addition to Group II decarboxylases. Proc Natl Acad Sci USA 108:20514-20519.

Glover J, Grelon M, Craig S, Chaudhury A, Dennis E (1998) Cloning and characterization of MS5 from *Arabidopsis*: a gene critical in male meiosis. Plant J 15:345-356.

Gonzalez-Carranza ZH, Elliott KA, Roberts JA (2007) Expression of polygalacturonases and evidence to support their role during cell separation processes in *Arabidopsis thaliana*. J Exp Bot 58:3719-3730.

Guillaumot D, Guillon S, Deplanque T, Vanhee C, Gumy C, Masquelier D, Morsomme P, Batoko H (2009) The *Arabidopsis* TSPO-related protein is a stress and abscisic acid-regulated, endoplasmic reticulum-Golgi-localized membrane protein. 60:242-256.

Guo M, Rupe MA, Dieter JA, Zou J, Spielbauer D, Duncan KE, Howard RJ, Hou Z, Simmons CR (2010) Cell Number Regulator1 affects plant and organ size in maize: implications for crop yield enhancement and heterosis. Plant Cell 22(4): 1057-73.

Greer S, Wen M, Bird D, Wu X, Samuels L, Kunst L, Jetter R (2007) The cytochrome P450 enzyme CYP96A15 is the midchain alkane hydroxylase responsible for the formation of secondary alcohols and ketones in stem cuticular wax of *Arabidopsis*. Plant Physiol. 145:653-667.

Guidotti BB, Gomes BR, Siqueira-Soares RDC, Soares AR, Ferrarese-Filho O (2013) The effects of dopamine on root growth and enzyme activity in soybean seedlings. Plant Signal Behav 8(9):e25477.

Guo S, Zhang J, Sun H et al. (2013) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nature Genetics 45:51-58.

Ma GS, Guo H, Jian C (2000) Study on the pattern of diseases in Chinese chestnuts in storage. Plant Protection 26(4):29-31.

Hayashi S, Ishii T, Matsunaga T, Tominaga R, Kuromori T, Wada T, Shinozaki K, Hirayama T (2008) The glyverophosphoryl diester phosphodiesterase-like proteins SHV3 and its homologs play important roles in cell wall organization. Plant Cell Physiol 49:1522-1535.

Huang HW, Norton JD, Smith DA, Slayden PW (1990) Characteristics of chestnut gall wasp resistant Chinese chestnuts. Annual Report of the Northern Nut Grower's Association 81:29-32.

Huang WG, Zhou ZJ, Cheng LL, Chen SF, He XS (2009) A new variety of Chinese chestnut 'Heishanzhai 7'. Scientia Silvae Sinicae 45(6):177.

Jackson JP, Lindroth AM, Cao X, Jacobsen SE (2002) Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature 416(6880):556-60.

Jiangsu Institute of Botanical Research (1979) Ban Li (Chestnut) Science Publishing House, Beijing, China. (In Chinese, partial translation available from The American Chestnut Foundation).

Jirschitzka J, Schmidt GW, Reichelt M, Schneider B, Gershenzon J, D'Auria JC (2012) Plant tropane alkaloid biosynthesis evolved independently in the Solanaceae and Erythroxylaceae. Proc Natl Acad Sci USA 109(26):10304-9.

Junhao D, Mingqing Z, Zongwen L, Zhongfu H (2000) Breeding research of Lantian bright red Chinese chestnut. Journal of Northwest Forestry College 1:004.

Kaga A, Isemura T, Tomooka N, Vaughan DA (2008) The genetics of domestication of the azuki bean (*Vigna angularis*) Genetics 178(2):1013-1036.

Khan MA, Olsen KM, Sover V, Kushad MM, Korban SS (2014) Fruit quality traits have played critical roles in domestication of the apple. The Plant Genome 7. doi:10.3835/plantgenome2014.04.0018

Kim SC, Kang L, Nagaraj S, Blancaflor EB, Mysore KS, Chapman KD (2009) Mutations in *Arabidopsis* fatty acid amide hydrolase reveal that catalytic activity influences growth but not sensitivity to abscisic acid or pathogens. J Biol Chem 284(49):34065-34074.

Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T (2001) Gene duplication in the diversification of secondary metabolism: tandem 2-oxyglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. Plant Cell 13:681-693.

Knip M, de Pater S, Hooykaas PJ (2012) The SLEEPER genes: a transposase-derived angiosperm-specific gene family. BMC Plant Biol 12:192-192.

Knoblauch M, Froelich DR, Pickard WF, Peters WS (2014) SEORious business: structural proteins in sieve tubes and their involvedment in sieve element occlusion. J Exp Bot 65(7):1879-1893.

Kofler R, Pandey PV, Schlötterer C (2011) PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinfomatics 27(24):3435-3436.

LalondeS, Sero A, Pratelli R et al. (2010) A membrane protein/signaling protein interaction network for *Arabidopsis* version AMPv2. Front. Physiol. 1:24-24.

Lam H-M, Xu X, Liu X, et al. (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature Genetics 42:1053:1059.

Lefebvre V, Kunst M, Camara B, Palloix A (1998) The capsanthin-capsorubin synthase gene: a candidate gene for the y locus controlling red fruit color in pepper. Plant Mol Biol 36:785-789.

Levin DA (1976) The chemical defenses of plants to pathogens and herbivores. Annual Review of Ecology and Systematics 7:121-159.

Levy YY, Mesnage S, Mylne JS, Gendall AR, Dean C (2002) Multiple roles of *Arabidopsis* VRN1 in vernalization and flowering time control. Science 297:243-246.

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science 311(5769):1936-1939.

Li HL, Willaman JJ (1968) Distribution of alkaloids in angiosperm phytogeny. Economic Botany 22(3):239-252.

Li X, Dodson J, Zhou X, Zhang H, Masutomoto R (2007) Early cultivated wheat and the broadening of agriculture in Neolithic China. The Holocene 17:555.

Li Y, Zhao S-C, Ma J-X et al. (2013) Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 14:579.

Liu L, Liu C, Hou X, Xi W, Shen L, Tao Z, Wang Y, Yu H (2012) FTIP1 is an essential regulator required for florigen transport. PLoS Biol 10:E1001313-E1001313.

Liu Y, Li X, Li K, Liu H, Lin C (2013) Multiple bHLH proteins form heterodimers to mediate CRY2-dependent regulation of flowering time in *Arabidopsis*. PLoS Genet 9:E1003861-E1003861.

Lukacin R, Wellmann F, Britsch L, Martens S, Matern U (2003) Flavonol synthase from *Citrus unshui* is a bifunctional dioxygenase. Phytochemistry 62:287-292.

Lynch M, Bost D, Wilson S, Maruki T, Harrison S (2014) Population-genetic inference from pooled-sequencing data. Genome Bio Evol. 6(5):1210-1218.

Mandoakar A, Browse J (2009) MYB108 acts together with MYB24 to regulate jasmonate-mediated stamen maturation in *Arabidopsis*. Plant Physiol. 149:851-862.

Mao L, Begum D, Chuang HW, Budiman MA, Szymkowiak EJ, Irish EE, Wing RA (2000) *Jointless* is a MADS-box gene controlling tomato flower abscission zone development. Nature 406:910-913.

Matsubayashi Y, Ogawa M, Kihara H, Niwa M, Sakagami Y (2006) Disruption and overexpression of *Arabidopsis* phytosulfokine receptor gene affects cellular longevity and potential for growth. Plant Physiol 142(1):45-53.

Mengiste T, Chen X, Salmeron J, Dietrich R (2003) The BOTRYTIS SUSCEPTIBLE1 gene encodes an R2RMYB transcription factor protein that is required for biotic and abiotic stress responses in *Arabidopsis*. Plant Cell 15:2551-2565.

Merkouropoulos G, Andreasson E, Hess D, Boller T, Peck SC (2008) An *Arabidopsis* protein phosphorylated in response to microbial elicitation, AtPHOS32, is a substrate of MAP kinases 3 and 6. J Biol Chem 283:10492-10499.

Metaxas A (2013) Chestnut (*Castanea* spp.) cultivar evaluation for commercial chestnut production in Hamilton County, Tennessee. M.S.E.S Thesis, University of Tennessee at Chattanooga, 135 pp.

Müller B, Fastner A, Karmann J et al. (2015) Amino acid export in developing *Arabidopsis* seeds depends on UmamiT facilitators. Current Biology 25:3126-3131.

Mueller D, Schmitz G, Theres K (2006) Blind homologous R2R3 Myb genes control the pattern of lateral meristem initiation in *Arabidopsis*. Plant Cell 18:586-597.

Naot D, Ben-Hayyim G, Eshdat Y, Holland D (1995) Drought, heat and salt stress induce the expression of a citrus homologue of an atypical late-embryogenesis Lea5 gene. Plant Mol Biol 27(3):619-22.

Nishio S, Yamada M, Takada N, Kato H, Onoue N, Sawamura Y, Saito T (2014) Environmental variance and broad-sense heritability of nut traits in Japanese chestnut breeding. HortScience 49(6):696-700.

Nito K, Kamigaki A, Kondo M, Hayashi M, Nishimura M (2007) Functional classification of *Arabidopsis* peroxisome biogenesis factors proposed from analysis of knockdown mutants. Plant Cell Physiol 48:763-774.

Ogawa M, Kay P, Wilson S, Swain SM (2009) *ARABIDOPSIS* DEHISCENCE ZONE POLYGALACTURONASE1 (ADPG1), ADPG2, and QUARTET2 are polygalacturonases required for cell separation during reproductive development in *Arabidopsis.* Plant Cell 21:216-233.

Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan (2006) Selection under domestication: evidence for a sweep in the rice *Waxy* genomic region. Genetics 173(2)975-983.

Ovesna J, Kucera L, Jiang LJ, Vagnerova D (2004) Characterisation of Chinese elite cultivars and genetic resources of chestnut by AFLP. Biologia Plantarum 49(1):125-127.

Paran I, van der Knaap E (2007) Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. Journal of Experimental Botany 58(14):3841-3852.

Pereira-Lorenzo S, Ramos-Cabrer AM, Ciordia-Ara M, Rios-Mesa D (2006) Chemical composition of chestnut cultivars from Spain. Scientia Horticulturae 9:134-42.

Pereira-Lorenzo S, Lourenço Costa R, Anagnostakis S et al. (2016) Interspecific hybridization of chestnut. In: Polyploidy and Hybridization for Crop Improvement, Chapter: 15, Publisher: CRC Press, Editor: Mason AS, pp. 379-408.

Piatkowski D, Schneider K, Salamini F, Bartels D (1990) Characterization of five abscisic acid-responsive cDNA clones isolated from the desiccation-tolerant plant *Craterostigma plantagineum* and their relationship to other water-stress genes. Plant Physiol. 94:1682-1688.

Preuss ML, Serna J, Falbel TG, Bednarek SY, Nielsen E (2004) The *Arabidopsis* Rab GTPase RABA4b localizes to the tips of growing root hair cells. Plant Cell 16:1589-1603.

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. Nature 457: 843-848.

Qi J, Liu X, Shen D et al. (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nature Genetics 45:1510-1515.

Qin C, Yu C, Shen Y et al. (2014) Whole-genome sequencing of cultivated and wild pepper provides insights into *Capsicum* domestication and specialization. Proc Nat Acad Sci USA 111(14):5135-5140.

Qin Y-M, Hu C-Y, Pang Y, Kastaniotis AJ, Hiltunen JK, Zhu Y-X (200&) Saturated very-long-chain fatty acids promote cotton fiber and *Arabidopsis* cell elongation by activating ethylene biosynthesis. Plant Cell 19:3692-3704.

Ramya KB, Thaakur S (2007) Herbs containing L-Dopa: an update. Anc Sci Lige 27(1):50-5.

Rao GU, Paran I (2003) Polygalacturonase: a candidate gene for the soft flesh and deciduous fruit mutation in *Capsicum.* Plant Mol Biol 51: 135-141.

Reumann S, Babujee L, Ma C, Wienkoop S, Siemsen T, Antonicelli GE, Rasche N, Lueder F, Weckwerth W, Jahn O (2007) Proteome analysis of *Arabidopsis* leaf peroxisomes reveals novel targeting peptides, metabolic pathways, and defense mechanisms. Plant Cell 19:3170-2193.

Ronen G, Carmel-Goren L, Zamir D, Hirschberg J (2000) An alternative pathway to beta-carotene formation in plant chloroplasts discovered by map-based cloning of *beta* and *old-gold* color mutations in tomato. Proc Nat Acad Sci USA 97:11102-11107.

Ross-Ibarra J, Morrell PL, Gaut BS (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Nat Acad Sci USA 104(1):1641-8648.

Rutter PA, Miller G, Payne JA (1991) Chestnuts (*Castanea*) Acta Hortic. 290:761-790.

Sasabe M, Toyoda K, Shiraishi T, Inagaki Y, Ichinose Y (2002) cDNA cloning and characterization of tobacco ABC transporter: NtPDR1 is a novel elicitor-responsive gene. FEBS let. 518:164-168.

Schnellmann S, Schnittger A, Kirik V, Wada T, Okada K, Beermann A, Thumfahrt J, Juergens G, Huelskamp M (2002) TRIPTYCHON and CAPRICE mediate lateral inhibition during trichome and rooth hair patterning in Arabidopsis. EMBO J 21:5036-5046.

Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, Yano M (2008) Deletion in a gene associated with grain size increased yields during rice domestication. Nature Genetics 40:1023-1029.

Sieber P, Gheyselinck J, Gross-Hardt R, Laux T, Grossnicklaus U, Schneitz K (2004) Pattern formation during early ovule development in *Arabidopsis thaliana*. Developmental Biology 273:321-334.

Sakai H, Medrano LJ, Meyerowitz EM (1995) Role of SUPERMAN in maintaining *Arabidopsis* floral whorl boundaries. Nature 378:199-203.

Sakuma S, Salomon B, Komatsuda T (2011) The domestication syndrome genes responsible for the major changes in plant form in the Triticeae crops. Plant Cell Physiol 52(5):738-749.

Schloetterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. Nat Rev Genet 15(11):749-63.

Schmutz J, McClean P, Mamdia S et al. (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics 46:707-713.

Shi Z, Stoesser R (2005) Reproductive biology of Chinese chestnut (*Castanea mollissima* Blume). Europ. J. Hort. Sci. 70(2):96-103.

Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai Y-S, Gill BS, Faris JD (2006) Molecular characterization of the major wheat domestication gene *Q*. Genetics 172(1):547-555.

Soederman E, Hjellstroem M, Fahleson J, Engstroem P (1999) 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 40:1073-1083.

Song W-Y, Wang G-L, Chen L-L et al. (1995) A receptor kinase-like protein encoded by the rice blast disease resistance gene, Xa21. Science 270:1804-1806.

Sprunck S, Rademacher S, Vogler F, Gheyselinck J, Grossniklaus U, Dresselhaus T (2012) Egg cell-secreted EC1 triggers sperm cell activation during double fertilization. Science 338:1093-1097.

Stanke M, Schoeffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62.

Staton ME, Addo-Quaye C, Cannon N, Tomsho LP, Drautz D, Wagner TK, Zembower N, Ficklin S, Saski C, Burhans R, Schuster SC, Abbott AG, Nelson CD, Hebard FV, Carlson JE (2014) *The Chinese chestnut (Castanea mollissima) genome version 1.1*, http://www.hardwoodgenomics.org/chinese-chestnut-genome, Access date August 2, 2016.

Syring JV, Tennessen JA, Jennings T, Wegrzyn J, Scelfo-Dalbey C, Cronn R (2016) Targeted capture sequencing in whitebark pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome. Front. Plant. Sci.7:484.

Takada N, Nishio S, Yamada M, Sawamura Y, Sato A, Hirabayashi T, Saito T (2012) Inheritance of the easy-peeling pellicle trait of Japanese chestnut cultivar Porotan. HortScience 47(7):845-847.

Tan L, Li X, Liu F, Sun X et al. (2008) Control of a key transition from prostrate to erect growth in rice domestication. Nature Genetics 40:1360-1364.

Tanimoto M, Tremblay R, Colasanti J (2008) Altered gravitropic response, amyloplast sedimentation, and circumnutation in the *Arabidopsis* shoot gravitropism 5 mutant are associated with reduced starch levels. Plant Mol Biol. 67:57-59.

Tsay YF, Chiu CC, Tsai CB, Ho CH, Hsu PK (2007) Nitrate transporters and peptide transporters. FEBS Lett. 581:2290-2300.

Uhrig RG, Moorhead GB (2011) Two ancient bacterial-like PPP family phosphatases from *Arabidopsis* are highly conserved plant proteins that possess unique properties. Plant Physiol 157:1778-1792.

Valerio L, De Meyer M, Penel C, Dunand C (2004) Expression analysis of the *Arabidopsis* peroxidase multigenic family. Phytochemistry 65(10):1331-42.

Velasco D, Hough J, Aradhya M, Ross-Ibarra J (2016) Evolutionary genomics of peach and almond domestication. G3 (Bethesda) 6(12):3985-3993.

Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, Drake R, Shuch W, Giovannoni J (2002) A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (*rin*) locus. Science 196:343-346.

Wang E, Wang J, Zhu X et al. (2008) Control of rice grain-filling and yield by a gene with a potential signature of domestication. Nature Genetics 40:1370-1374.

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bomblies K, Lukens L, Doebley JF (2005) The origin of the naked grains of maize. Nature 436:714-719.

Wang H, Zhu Y, Fujioka S, Asami T, Li J, Li J (2009) Regulation of *Arabidopsis* brassinosteroid signaling by atypical basic helix-loop-helix proteins. Plant Cell 21:3781-3791.

Yamagami T, Tsuchisaka A, Yamada K, Haddon WF, Harden LA, Theologis A (2003) Biochemical diversity among the 1-amino-cyclopropane-carboxylate synthase isozymes encoded by the *Arabidopsis* gene family. J Biol Chem 278:49102-49112.

Yanofsky MF, Ma H, Bowman JL, Drews G, Feldmann KA, Meyerowitz EM (1990) The protein encoded by the *Arabidopsis* homeotic gene agamous resembles transcription factors. Nature 346:35-39.

Xi E-H, Zhang H-B, Sheng J et al. (2017) The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Molecular Plant 10(6):866-877.

Xu YH, Jiang YC, Wang ZJ, Fang B, Wang QF, Zhang LT, Su XG, Wang CH (2010) Breeding of a new processing Chinese chestnut cultivar Jinliwang. J Fruit Sci 27(1):156-157.

Yang F, Liu Q, Pan S, Xu C, Youling L, Xiong L (2015) Chemical composition and quality traits of Chinese chestnuts (Castanea mollissima) produced in different ecological regions. Food BioScience 11:33-42.

Youngstaedt E, Gurney K (2013) Chestnut growers guide to pests and diseases. Journal of the American Chestnut Foundation 27(3):17-23.

Yu Y, Tang T, Qian Q et al. (2008) Independent losses of function in a polyphenol oxidase in rice: differentiation in grain discoloration between subspecies and the role of positive selection in domestication. The Plant Cell 20(11):2946-2959.

Zhang YL, Shao ZX, Yang WM, Ning DL, Du CH (2010) Selection of a new Chinese chestnut cultivar Yunxia. J Fruit Sci 27(3):475-476.

Zhou Y, Zhu J, Li Z, Yi C, Liu J, Zhang H, Tang S, Gu M, Liang G (2009) Deletion in a quantitative trait gene qPE9-1 associated with panicle erectness improves plant architecture during rice domestication. Genetics 183(1):315-324.

Zohary D (2004) Unconcious selection and the evolution of domesticated plants. Economic Botany 58(1):5-10.

Zorzatto C, Machado JP, Lopes KV et al. (2015) NIK1-mediated translation suppression functions as a plant antiviral immunity mechanism. Nature 520:679-682.

Table 5.1. *Castanea mollissima* DNA samples pools, with individuals (n) per sample site.

Pool	Location	Origin	n	Bases (quality clipped)	Estimated depth
Y1	Yunnan- 26.013° N 101.0932° E	Forest	9	7,375,355,857	9.46
Y2	Yunnan- Fengqing County	Forest	10	11,619,019,171	14.89
S1	Shaanxi- Zhuque and Heihe Forests	Forest	13	7,344,176,568	9.42
S2	Shaanxi- Ningshan County	Forest	10	8,630,350,892	11.06
S3	Shaanxi- Liuba County	Forest	10	5,612,880,521	7.19
S4	Shaanxi-33.772° N 108.766° E	Orchard	10	9,890,960,153	12.68
GZ	Guizhou- 26.236° N 105.1676° E	Forest	10	8,594,920,262	11.02
HB	Hebei- 40.597° N 118.399° E	Orchard	10	6,175,704,628	7.92
BY	Beijing- Yanqing County	Orchard	10	5,240,552,948	6.72
ECC	Ohio, U.S.A.[a]	Orchard	12	3,498,422,441	4.49

[a] Grown in Ohio, USA; derived from northern Chinese orchard cultivars

Table 5.2. Summary of SNP calls in each sample pool across the entire *C. mollissima* genome.

Pool	Variant sites[a]	Sites with alt >0.2[b]	Average Depth[c]
Y1	8564349	6359398	12.38
Y2	11431073	9371773	17.56
S1	13492467	8279766	9.46
S2	7632501	5074304	9.12
S3	10272008	8155228	13.09
S4	17086140	12681410	13.96
GZ	10046443	7949407	14.10
HB	7796678	5479063	11.03
BY	6844594	4301595	9.03
ECC	4939429	2079207	7.10

[a] Sites with a variant called in a given pool with read depth >6; [b] Sites with an alternate allele of frequency at least 0.2; [c] Average depth at variant sites numbered in the 2nd column.

Table 5.3 Putative selection intervals identified by lower values of Tajima's D and π in the orchard pool vs. the wild pool, with annotations based on UniProt alignments of predicted genes.

LG[a]	Start[b]	End	D_o[c]	D_w[d]	p[e]	Predicted Function	Aligned Gene
LGA	9120000	9150000	-2.01	-0.09	0.0049	Uncharacterized protein Y1491	ACCO1_ARATH
LGA	17560000	17720000	-1.57	0.31	0.0060	1-aminocyclopropane-1-carboxylate oxidase	ACCO4_ARATH
LGA	25700000	25770000	-1.85	0.67	<0.001	NRT/PTR family protein PTR9	PTR9_ARATH
LGA	28940000	28950000	-1.65	1.23	<0.001	Putative phytosulfokines PSK6	PSK6_ARATH
LGA	30430000	30480000	-1.59	0.42	0.003	Methyltransferase-like protein 16	MET16_HUMAN
LGA	39980000	40020000	-1.79	0.20	0.003	Glycerophosphodiester phosphodiesterase	GPDL3_ARATH
LGA	46000000	46010000	-2.03	-0.47	0.020	Flowering time control protein FCA	FCA_ARATH
LGA	46360000	46410000	-1.98	-0.05	0.005	Alkane hydroxylase MAH1	MAH1_ARATH
LGA	46500000	46540000	-1.76	-0.09	0.014	Transcription factor MYB108	MY108_ARATH
LGA	48500000	48570000	-1.49	0.29	0.009	2-oxyglutarate-dependent dioxygenase	AOP1C_ARATH
LGA	53300000	53320000	-1.25	0.44	0.013	Transcription factor bHLH78	BH078_ARATH
LGA	53710000	53750000	-1.70	-0.07	0.016	Late embryogenesis abundant protein	LEA5_CITSI
LGA	58690000	58730000	-2.12	-0.47	0.015	Probable polygalacturonase	ADPG2_ARATH
LGA	66030000	66050000	-1.92	0.31	<0.001	Vesicle transport v-SNARE 13	VTI13_ARATH
LGA	72490000	72550000	-1.59	0.07	0.019	Anthocyanidin 3-O-glucosyltransferase 2	UFOG_VITVI
LGA	87260000	87310000	-2.14	0.10	<0.001	Translocator protein homolog	TSPO_ARATH
LGA	103940000	104000000	-2.10	0.19	<0.001	Importin subunit alpha-2	IMPA2_ARATH
LGA	104070000	104130000	-1.86	0.13	0.003	Transcription factor GATA-type	GAT4_ARATH
LGB	3070000	3180000	-1.92	-0.08	0.011	Homeobox-leucine zipper protein	ATHB6_ARATH
LGB	6610000	6660000	-2.06	0.29	<0.001	Sieve-element occlusion protein	SEOB_ARATH
LGB	7550000	7610000	-1.87	-0.03	0.011	Probable enoyl-CoA hydratase 1	ECH1P_ARATH
LGB	7750000	7880000	-1.64	0.43	0.004	Transcription factor bHLH147	BH147_ARATH

Table 5.3 continued

LG	Start	End	D_o	D_w	p	Predicted Functions	Best Gene Alignment
LGB	8100000	8180000	-2.51	0.17	<0.001	Desiccation-related protein PCC13-62	DRPE_CRAPL
LGB	8690000	8870000	-1.71	0.17	0.009	Zinc finger BED domain protein RICESLEEPER 1	RSLE1_ORYSJ
LGB	14200000	14240000	-1.67	0.36	0.005	G-type lectin S-receptor kinase	Y1130_ARATH
LGB	19610000	19740000	-1.96	0.44	<0.001	Sucrose synthase 2	SUS2_ARATH
LGB	20350000	20370000	-1.57	0.04	0.022	Homeobox-leucine zipper protein	ATB14_ARATH
LGB	20540000	20580000	-1.76	0.10	0.009	LanC-like protein	GCL1_ARATH
LGB	31040000	31160000	-1.70	0.19	0.009	Geraniol 8-hydroxylase C76C3	C76B6_CATRO
LGB	31310000	31350000	-1.74	-0.27	0.015	BI1-like protein	BI1L_ARATH
LGC	5470000	5470000	-1.78	0.51	<0.001	Pleiotropic drug resistance protein 1	PDR1_TOBAC
LGC	6500000	6500000	-1.99	0.29	<0.001	Cytochrome P450 CYP71D312	C7D31_PANGI
LGC	19650000	19650000	-1.76	-0.06	0.013	Heat shock factor protein	HSFB1_ARATH
LGC	21310000	21310000	-1.65	0.07	0.013	Reticuline oxidase	RETO_PAPSO
LGC	25700000	25700000	-1.97	0.34	<0.001	Peroxidase 29	PER29_ARATH
LGC	30050000	30050000	-1.68	0.56	<0.001	Acyl carrier protein 5	ACP5_ARATH
LGC	30360000	30360000	-1.85	0.43	<0.001	Histone-lysine N-methyltransferase SUVH4	SUVH4_ARATH
LGC	49350000	49350000	-2.15	-0.11	0.003	1-aminocyclopropane-1-carboxylate oxidase 4	ACCO4_ARATH
LGC	49920000	49920000	-1.61	0.27	0.007	F-box protein SKIP28	SKI28_ARATH
LGC	50050000	50050000	-1.52	0.15	0.015	Cell number regulator 2	CNR2_MAIZE
LGC	50850000	50850000	-2.18	-0.34	0.008	Transcription factor RAX2	RAX2_ARATH
LGD	4350000	4430000	-1.53	0.49	0.003	NADH dehydrogenase [ubiquinone] Fe-S protein	NDUS3_ARATH
LGD	5910000	5930000	-1.51	0.09	0.018	Egg cell-secreted protein 1.1	EC11_ARATH
LGD	7620000	7730000	-1.70	0.15	0.009	SHOOT GRAVITROPISM 5	IDD15_ARATH
LGD	14190000	14290000	-1.44	0.33	0.009	RING-H2 finger protein ATL54, flavonol synthase	ATL54_ARATH, FLS_SITUB
LGD	18020000	18600000	-1.94	-0.13	0.007	Major allergens Pru1, Mal11	PRU1_PRUAR

Table 5.3 continued

LG	Start	End	D_o	D_w	p	Predicted Functions	Best Gene Alignment
LGD	32730000	32890000	-1.29	0.52	0.007	Transcription factor Scarecrow-like protein 30	SCL30_ARATH
LGD	34040000	34070000	-1.37	0.27	0.015	Aromatic-L-amino-acid decarboxylase	DDC_HUMAN
LGD	34630000	34680000	-1.46	0.46	0.005	G-type lectin S-receptor-like kinase	Y4230_ARATH
LGD	34780000	34790000	-1.86	0.13	0.003	G-type lectin S-receptor-like kinase	Y4230_ARATH
LGD	53450000	59960000	-1.16	0.41	0.020	Transcriptional regulator TIFY 6B	TIF6B_ARATH
LGD	60340000	60370000	-1.39	0.29	0.012	WAT1-related protein At3g53210	WTR26_ARATH
LGE	4980000	5000000	-1.70	0.40	<0.001	Fatty acid amide hydrolase	FAAH_ARATH
LGE	6140000	6210000	-2.16	-0.16	0.003	Polygalacturonase At1g48100	PGLR4_ARATH
LGE	6300000	6410000	-2.04	-0.01	0.003	Signal recognition particle 19 kDa protein	SRP19_ARATH
LGE	13940000	13990000	-1.88	0.07	0.004	ABIL2 actin-regulatory protein	ABIL2_ARATH
LGE	15260000	15340000	-2.24	-0.55	0.013	Tropinone reductase-like 3	TRPL3_ERYCB
LGE	16710000	16820000	-1.07	0.69	0.009	Lamin-like protein LAML	LAML_ARATH
LGE	19970000	20010000	-1.79	0.25	0.002	Exopolygalacturonase	Q39094_ARATH
LGE	20090000	21160000	-1.77	0.02	0.009	Shewanella-like protein phosphatase 2	SLP2_ARATH
LGE	30630000	30680000	-1.55	0.28	0.007	Glucan endo-1,3-beta-glucosidase 6	E1310_ARATH
LGE	42970000	43090000	-1.62	0.21	0.007	PLASTID TRANSCRIPTIONALLY ACTIVE 12	PTA12_ARATH
LGE	50780000	50810000	-1.62	0.08	0.012	POLLENLESS 3	MS5_ARATH
LGF	7900000	7980000	-1.71	0.39	0.001	G-type lectin S-receptor kinase RKS1	RKS1_ARATH
LGG	6870000	6940000	-1.81	0.18	0.003	G-type lectin S-receptor kinase Y1135	Y1135_ARATH
LGG	13830000	13850000	-1.53	0.39	0.004	Peroxisome biogenesis protein 1	PEX1_ARATH
LGG	24830000	24860000	-1.87	0.09	0.003	Late embryogenesis abundant protein D-29	LEA29_GOSHI

Table 5.3 (continued)

LG[a]	Start[b]	End	D_o	D_w	p[c]	Predicted Functions	Best Gene Alignment
LGG	49050000	49110000	-2.08	-0.29	0.008	Syntaxin-132, isocitrate lyase	SY132_ARATH, ACEA_GOSHI
LGH	780000	820000	-2.23	0.35	<0.001	Ras-related protein RAA4b	RAA4B_ARATH
LGH	38320000	38370000	-2.08	0.17	0.003	Probable LRR receptor-like kinase At2g28990	Y2899_ARATH
LGH	40870000	40890000	-1.99	-0.23	0.021	BAG family molecular chaperone regulator 4	BAG4_ARATH
LGH	40920000	40930000	-1.83	0.00	0.016	L10-interacting MYB domain protein	LIMYB_ARATH
LGI	10470000	10560000	-1.59	0.34	0.008	Syntaxin-132	SY132_ARATH
LGI	33370000	33450000	-2.08	0.33	<0.001	FT (Flowering Time) - interacting protein 1	FTIP1_ARATH
LGI	40500000	40580000	-1.94	0.33	0.001	Peroxidase 24	PER24_ARATH
LGJ	1950000	2040000	-2.18	-0.02	0.002	Topless-related protein 2	TPR2_ARATH
LGJ	16440000	16490000	-2.09	0.03	0.002	Transcription factor VRN1	VRN1_ARATH
LGJ	16890000	16990000	-1.78	0.28	0.003	Universal stress protein PHOS34	PHOS34_ARATH
LGJ	22590000	22650000	-1.92	0.25	0.002	Major allergen Pru1	PRU1_PRUAR
LGJ	47610000	47750000	-1.98	0.12	0.003	Transcription factor ERF3	ERF82_ARATH
LGK	19140000	19230000	-1.84	0.22	0.002	Receptor kinase-like protein Xa21	XA21_ORYSJ
LGK	25200000	25270000	-2.02	0.10	0.001	Floral homeotic protein AGAMOUS	AG_ARATH
LGL	18850000	18890000	-1.80	-0.19	0.016	Actin 1, protein kinase inhibitor SMR3	ACT1_ARATH
LGL	25960000	25980000	-1.64	-0.10	0.021	Multidrug resistance protein	AB19B_ARATH
LGL	37920000	38000000	-2.07	0.16	<0.001	Peroxidase 16	PER16_ARATH
LGL	38090000	38170000	-2.08	0.25	<0.001	1-aminocyclopropane-1-carboxylate synthase 7	ACS7_ARATH

[a]Pseudochromosome reference sequence corresponding to the given linkage group; [b]bp position at the start of the interval where a significant difference was detected between orchard tree and wild tree samples based on the Chinese chestnut reference genome draft assembly (Carlson et al.); [c]p-value based on 5000 permutations of the whole-genome dataset, [d]Predicted function based on Uniprot alignments to predicted genes within the specified interval; [e]Uniprot identifier for selected genes within the specified region

Table 5.4 Putative domestication loci identified by comparing allele frequencies among wild and domestic pools of chestnut, with annotations based on the best UniProt alignments of predicted genes.

LG	Start	End	O[a]	W[b]	Sd[c]	Predicted genes
LGA	50030000	50070000	0.93	0.63	>3	SMH4_MAIZE single myb histone 4, TRB4_ARATH telomere repeat-binding factor, THOC6_ARATH THO abscisic acid signalling regulator
LGA	65859000	65955000	0.86	0.56	>3	RPE_SOLTU ribulose-phosphate 3-epimerase
LGA	89343000	89450000	0.72	0.59	>3	AUR3_ARATH Aurora-3 serine-threonine protein kinase
LGB	25835000	25936000	0.86	0.64	>3	P2C78_ORYSJ probable protein phosphatase
LGB	40045000	40092000	0.88	0.57	>3	DTX27_ARATH efflux transporter
LGC	31478000	31571000	0.93	0.68	>3	PP14_ARATH Serine/threonine-protein phosphatase PP1
LGE	12050000	12158000	0.96	0.71	>3	ORG2_ARATH transcription factor
LGE	25481000	25588000	0.87	0.71	>3	MYBF_ARATH transcription factor, E131_ARATH Glucan endo-1,3-beta-glucosidase 1
LGE	29262000	29326000	0.93	0.66	>3	CE101_ARATH G-type lectin S-receptor-like serine/threonine-protein kinase, promotes expression of photosynthesis-related genes
LGE	34936000	34951000	0.87	0.67	>3	UBP13_ARATH ubiquitin carboxyl-terminal hydrolase 13
LGE	44249000	44394000	0.92	0.61	>5	DRE2D_ARATH dehydration-responsive element, WAT1_ARATH WALLS ARE THIN 1 indole metabolism protein, VHAA2_ARATH V-type ATPase subunit a2, ZOG_PHALU zeatin O-glucosyltransferase
LGF	16717000	16833000	0.84	0.60	>3	PERK9_ARATH proline-rich receptor-like protein kinase, RABA3_ARATH Ras-related protein (protein transport), AL3F1_ARATH aldehyde dehydrogenase family 3 member
LGF	27956000	18025000	0.84	0.63	>3	ATB12_ARATH homeobox-leucine zipper protein, SAU32_ARATH auxin-responsive protein SAUR32, CAMK3_ARATH CDPK-related kinase
LGG	8420000	8480000	0.96	0.64	>3	CDPK_SOYBN calcium-dependent protein kinase SK5
LGG	2054000	20710000	0.90	0.68	>3	GONS1_ARATH GDP-mannose transporter
LGG	23410000	23452000	0.85	0.59	>3	LEA34_GOSHI late embryogenesis-abundant protein, OFP12_ARATH transcriptional repressor that regulates BLH and KNAT transcription factors
LGG	33377000	33518000	0.92	0.64	>3	MEE14_ARATH CCG-binding protein, interacts with AGAMOUS-like transcription factors
LGH	35413000	35507000	0.89	0.63	>3	FB37_ARATH F-box ubiquitin-protein transeferase, RD21A_ARATH cysteine proteinase
LGK	26487000	26659000	0.86	0.68	>4	FRO6_ARATH ferric reduction oxidase, GUN25_ARATH endoglucanase 25

LG = Linkage group; Start and End locations based on draft reference genome assembly (Carlson et al.) [a]Average major allele frequency in orchard trees for the given interval of 10 predicted genes; [b]Average major allele frequency in wild trees for the given interval of

10 predicted genes; [c]Standard deviations greater than the average difference in major allele frequency between orchard tree and wild tree pools.

Table 5.5 Putative loci differentially selected among northern and southern samples of wild Chinese chestnut, identified by comparing allele frequencies among pools of chestnut, with annotations based on the best UniProt alignments of predicted genes.

LG	Start	End	N[a]	S[b]	Sd[c]	Predicted gene
LGA	72907000	72988000	0.56	0.87	>2	C94A2_VICSA: cytochrome P450, fatty acid oxidation
LGA	79800000	79880000	0.67	0.95	>3	Y2060_ARATH: BTB/POZ domain ubiquination protein
LGA	80300000	80330000	0.61	0.89	>3	SD25_ARATH: protein kinase
LGA	82239000	82355000	0.65	0.99	>4	PLY19_ARATH: pectate lyase 19
LGB	15342000	15410000	0.64	0.91	>3	E134_MAIZE: endo-1,3;1,4-beta-D-glucanase
LGB*	6540000	6678000	0.71	0.95	>3	SEOB_ARATH*, sieve-element occlusion protein
LGC	48510000	48811632	0.66	0.83	>2	SIB1_ARATH: sigma binding factor, pathogen defense
LGC*	50000000	50186000	0.69	0.95	>3	CNR2_MAIZE*, cell-number regulator
LGC	53870000	53947000	0.57	0.86	>3	PP413_ARATH: pentatricopeptide repeat-containing protein
LGE	16600000	16700000	0.67	1.00	>4	CCR1_ARATH: cinnamoyl-CoA reductase, lignin synthesis
LGG	43890000	43990000	0.70	0.99	>3	HMDH1_GOSHI: isprenoid precursor (mevalonate) synthesis
LGG	48970000	49040000	0.61	0.88	>3	CPC_ARATH: trichome development transcription factor
LGI	4295000	4336000	0.62	0.89	>4	SILD_FORIN, ILR1_ARATH: lignin biosynthesis
LGL	58890000	59190000	0.69	0.97	>4	ERF25_ARATH, DRE1B_ARATH: cold tolerance
LGE	34000000	34100000	0.91	0.65	>3	CADH_EUCBO: cinnamoyl alcohol dehydrogenase, lignin synth.
LGH	18500000	18547000	0.81	0.62	>3	SAG13_ARATH: senescence-associated protein

[a]Average major allele frequency in northern Chinese wild trees for the given interval of 10 predicted genes; [b]Average major allele frequency in southern Chinese wild trees for the given interval of 10 predicted genes; [c]Standard deviations greater than the average difference in major allele frequency between orchard and wild pools.

[*]Also identified as putative domestication loci due to low Tajima's D value in orchard samples.

Table 5.6 Summary of all putative domestication loci, with heterozygosity and pi for *Castanea mollissima* and *Castanea dentata* calculated for the most likely candidate predicted gene in each interval, and F_{ST} between species and inferred function of the predicted gene depicted. Boldface indicates loci with the strongest evidence.

LG	Start	Gene name[a]	Inferred function[b]	het_cm[c]	pi-cm[d]	pi-cd[e]	cd/cm[f]	F_{ST}[g]
LGA	9120000	lga_g1087	Uncharacterized protein Y1491	0.11	0.0014	0.0021	1.53	0.81
LGA	17560000	lga_g2116	1-aminocyclopropane-1-carboxylate oxidase	0.19	0.0014	0.0015	1.11	0.76
LGA	25700000	lga_g3150	NTR/PTR family protein PTR9	0.12	0.0015	0.0029	2.00	0.85
LGA	28940000	lga_g3565	Putative phytosulfokies PSK6	0.14	0.0009	0.0037	4.26	0.44
LGA	30430000	lga_g3757	Methyltransferase-like protein 16	0.11	0.0006	0.0017	2.80	0.36
LGA	39980000	lga_g5028	Glycerophosphodiester phosphodiesterase	0.29	0.0016	0.0005	0.29	0.61
LGA	**46000000**	**lga_g5798**	**Flowering time control protein FCA**	**0.00**	**0.0000**	**0.0005**	**483.33**	**0.81**
LGA	46360000	lga_g5850	Alkane hydroxylase MAH1	0.05	0.0013	0.0032	2.54	0.92
LGA	46500000	lga_g5869	Transcription factor MYB108	0.01	0.0003	0.0000	0.03	0.87
LGA	48500000	lga_g6111	2-oxyglutarate-dependent dioxygenase AOP1	0.05	0.0008	0.0006	0.83	0.93
LGA	50030000	lga_g6327	THO complex subunit, ABA signalling	0.05	0.0004	0.0019	4.47	0.82
LGA	53300000	lga_g6764	Transcription factor bHLH78	0.11	0.0009	0.0016	1.73	0.82
LGA	**53710000**	**lga_g6816**	**Late embryogenesis abundant protein LEA5**	**0.10**	**0.0001**	**0.0007**	**4.56**	**0.85**
LGA	58690000	lga_g7476	Probable polygalacturonase ADPG2	0.10	0.0008	0.0014	1.66	0.64
LGA	65859000	lga_g8373	Ribulose-phosphate epimerase	0.04	0.0006	0.0021	3.53	0.93
LGA	66030000	lga_g8383	Vesicle transport v-SNARE 13	0.10	0.0005	0.0010	1.87	0.63
LGA	72490000	lga_g9205	Anthocyanidin 3-O-glucosyltransferase 2	0.10	0.0008	0.0017	2.08	0.89
LGA	87260000	lga_g11094	Translocator protein homolog TSPO	0.08	0.0006	0.0018	2.79	0.85
LGA	89343000	lga_g11363	AURORA protein kinase	0.13	0.0006	0.0014	2.48	0.69
LGA	**103940000**	**lga_g13067**	**Importin subunit alpha-2**	**0.06**	**0.0005**	**0.0024**	**4.69**	**0.87**
LGA	104070000	lga_g13074	Transcription factor GATA-type	0.07	0.0012	0.0022	1.83	0.86

Table 5.6 continued

LG	Start	Gene name[a]	Inferred function[b]	het_cm[c]	pi-cm[d]	pi-cd[e]	cd/cm[f]	F_{ST}[g]
LGB	3070000	lgb_g404	Homeobox-leucine zipper protein ATHB-14	0.12	0.0009	0.0012	1.30	0.74
LGB	6610000	lgb_g846	Sieve-element occlusion protein SEOB	0.10	0.0005	0.0007	1.49	0.82
LGB	7550000	lgb_g975	Probable enoyl-CoA hydratase 1, peroxisomal	0.15	0.0009	0.0012	1.29	0.57
LGB	7750000	lgb_g1007	Transcription factor bHLH147	0.14	0.0031	0.0043	1.37	0.38
LGB	8100000	lgb_g1054	Desiccation-related protein PCC13-62	0.12	0.0030	0.0061	2.05	0.70
LGB	8690000	lgb_g1139	ZF-BED domain protein RICESLEEPER 1	0.20	0.0012	0.0018	1.50	0.81
LGB	14200000	lgb_g1844	G-type lectin S-receptor kinase Y1130	0.20	0.0044	0.0064	1.43	0.73
LGB	19610000	lgb_g2523	Sucrose synthase 2	0.16	0.0022	0.0045	2.04	0.46
LGB	20350000	lgb_g2608	Homeobox-leucine zipper protein ATHB-6	0.08	0.0021	0.0023	1.12	0.77
LGB	20540000	lgb_g2625	LanC-like protein GCL1	0.10	0.0006	0.0015	2.49	0.83
LGB	25835000	lgb_g3303	Serine/threonine protein phosphatase	0.10	0.0010	0.0021	1.99	0.85
LGB	31040000	lgb_g3941	Geraniol 8-hydroxylase C76C3	0.09	0.0012	0.0025	2.07	0.82
LGB	31310000	lgb_g3977	BI1-like protein BI1L	0.11	0.0020	0.0050	2.49	0.74
LGB	40045000	lgb_g5010	DTX27_ARATH, Y091_NPVOP	0.16	0.0009	0.0029	3.28	0.79
LGC	5470000	lgc_g661	Pleiotropic drug resistance protein 1	0.11	0.0006	0.0011	1.92	0.70
LGC	**6500000**	**lgc_g807**	**Cytochrome P450 CYP71D312**	**0.07**	**0.0019**	**0.0125**	**6.61**	**0.82**
LGC	19650000	lgc_g2396	Heat shock factor protein HSFB1	0.17	0.0024	0.0040	1.67	0.65
LGC	21310000	lgc_g2594	Reticuline oxidase	0.07	0.0008	0.0024	2.84	0.86
LGC	25700000	lgc_g3175	Peroxidase 29	0.14	0.0003	0.0005	1.65	0.86
LGC	30050000	lgc_g3763	Acyl carrier protein 5	0.36	0.0024	0.0003	0.14	0.37
LGC	30360000	lgc_g3816	Histone-lysine N-methyltransferase SUVH4	0.23	0.0012	0.0008	0.72	0.68
LGC	31478000	lgc_g3960	Serine/threonine protein phosphatase	0.04	0.0006	0.0019	3.27	0.72
LGC	**49350000**	**lgc_g6157**	**1-aminocyclopropane-1-carboxylate oxidase 4**	**0.04**	**0.0001**	**0.0017**	**17.21**	**0.91**
LGC	50050000	lgc_g6227	Cell number regulator 2	0.22	0.0017	0.0016	0.94	0.59
LGC	**50850000**	**lgc_g6330**	**Transcription factor RAX2**	**0.02**	**0.0005**	**0.0029**	**5.90**	**0.97**
LGD	4350000	lgd_g575	NADH dehydrogenase Fe-S protein	0.09	0.0005	0.0017	3.67	0.83
LGD	5910000	lgd_g783	Egg cell-secreted protein 1.1	0.06	0.0011	0.0011	1.01	0.88

Table 5.6 continued

LG	Start	Gene name[a]	Inferred function[b]	het_cm[c]	pi-cm[d]	pi-cd[e]	cd/cm[f]	F_{ST}[g]
LGD	7620000	lgd_g1017	SHOOT GRAVITROPISM 5, IDD15	0.10	0.0002	0.0008	3.34	0.58
LGD	14190000	lgd_g1823	RING-H2 finger protein ATL54	0.15	0.0029	0.0057	1.98	0.46
LGD	**18020000**	**lgd_g2376**	**Major allergens Pru1**	**0.12**	**0.0006**	**0.0033**	**5.34**	**0.88**
LGD	32730000	lgd_g4271	Transcription factor Scarecrow-like protein 30	0.13	0.0016	0.0011	0.69	0.80
LGD	34040000	lgd_g4440	UCRIA_PEA cytochrome B6-f complex	0.17	0.0035	0.0107	3.03	0.56
LGD	34630000	lgd_g4494	G-type lectin S-receptor-like kinase Y4230	0.21	0.0026	0.0037	1.44	0.48
LGD	34780000	lgd_g4511	G-type lectin S-receptor-like kinase Y4230	0.11	0.0016	0.0027	1.68	0.80
LGD	53450000	lgd_g7746	Transcriptional regulator TIFY 6B	0.16	0.0009	0.0007	0.82	0.60
LGD	60340000	lgd_g7797	WAT1-related protein At3g53210	0.17	0.0003	0.0000	0.00	nc
LGE	4980000	lge_g5194	Fatty acid amide hydrolase	0.11	0.0005	0.0007	1.34	0.62
LGE	6140000	lge_g797	Polygalacturonase At1g48100	0.36	0.0014	0.0011	0.82	0.58
LGE	6300000	lge_g819	Signal recognition particle 19 kDa protein	0.24	0.0021	0.0026	1.25	0.54
LGE	**12050000**	**lge_g1551**	**Transcription factor ORG2**	**0.07**	**0.0004**	**0.0021**	**4.79**	**0.62**
LGE	13940000	lge_g1767	WAVE complex protein ABIL2	0.05	0.0007	0.0014	1.99	0.73
LGE	15260000	lge_g1945	Tropinone reductase-like 3	0.08	0.0007	0.0012	1.69	0.73
LGE	16710000	lge_g2142	Lamin-like protein LAML	0.02	0.0004	0.0010	2.19	0.88
LGE	19970000	lge_g2533	Exopolygalacturonase	0.48	0.0038	0.0030	0.80	0.39
LGE	20090000	lge_g2540	Shewanella-like protein phosphatase 2	0.29	0.0049	0.0021	0.42	0.65
LGE	**25481000**	**lge_g3428**	**MYBF_ARATH, E131_ARATH**	**0.08**	**0.0007**	**0.0020**	**2.87**	**0.67**
LGE	**29262000**	**lge_g3727**	**Lectin receptor kinase CES101**	**0.07**	**0.0003**	**0.0057**	**18.48**	**0.82**
LGE	30630000	lge_g3906	Glucan endo-1,3-beta-glucosidase 10	0.24	0.0011	0.0015	1.34	0.79
LGE	34936000	lge_g4431	WEB family protein	0.08	0.0004	0.0029	6.99	0.79
LGE	42970000	lge_g5457	PLASTID TRANSCRIPTIONALLY ACTIVE	0.05	0.0010	0.0016	1.53	0.94
LGE	**44249000**	**lge_g5605**	**Dehydration-responsive element**	**0.03**	**0.0002**	**0.0015**	**6.67**	**0.85**
LGE	**50780000**	**lge_g6427**	**POLLENLESS 3, MS5**	**0.04**	**0.0014**	**0.0116**	**8.25**	**0.63**
LGF	7900000	lgf_g934	G-type lectin S-receptor kinase RKS1	0.06	0.0020	0.0086	4.30	0.67

Table 5.6 continued

LG	Start	Gene name[a]	Inferred function[b]	het_cm[c]	pi-cm[d]	pi-cd[e]	cd/cm[f]	F_{ST}[g]
LGF	16717000	lgf_g2053	Aldehyde dehydrogenase family protein	0.09	0.0002	0.0009	5.13	0.86
LGF	27956000	lgf_g3433	SAUR32 Auxin responsive element	0.09	0.0004	0.0007	1.81	0.90
LGG	2054000	lgg_g2558	GDP-mannose transporter	0.08	0.0006	0.0016	2.63	0.86
LGG	**6870000**	**lgg_g872**	**G-type lectin S-receptor kinase Y1135**	**0.06**	**0.0008**	**0.0049**	**5.81**	**0.81**
LGG	8420000	lgg_g5871	CDPK_SOYBN	0.11	0.0015	0.0025	1.64	0.49
LGG	**13830000**	**lgg_g2955**	**Peroxisome biogenesis protein 1**	**0.00**	**0.0001**	**0.0027**	**23.02**	**0.94**
LGG	**23410000**	**lgg_g1699**	**Late embryogenesis abundant protein**	**0.20**	**0.0004**	**0.0015**	**3.82**	**0.68**
LGG	33377000	lgg_g4266	MEE14_ARATH	0.11	0.0009	0.0029	3.16	0.78
LGG	49050000	lgg_g6369	Syntaxin-132, isocitrate lyase	0.13	0.0008	0.0015	1.84	0.73
LGH	780000	lgh_g105	Ras-related protein RAA4b	0.17	0.0025	0.0046	1.88	0.67
LGH	**35413000**	**lgh_g4460**	**F-box protein**	**0.05**	**0.0007**	**0.0025**	**3.88**	**0.84**
LGH	38320000	lgh_g4828	Probable LRR receptor-like kinase At2g28990	0.11	0.0010	0.0016	1.54	0.75
LGH	40870000	lgh_g5136	BAG family molecular chaperone regulator 4	0.12	0.0017	0.0037	2.20	0.70
LGH	40920000	lgh_g5143	L10-interacting MYB domain protein	0.36	0.0058	0.0013	0.23	0.54
LGI	10470000	lgi_g1363	Syntaxin-132	0.19	0.0010	0.0006	0.58	0.68
LGI	33370000	lgi_g4214	FT (Flowering Time) -interacting protein 1	0.10	0.0017	0.0023	1.36	0.77
LGI	40500000	lgi_g5153	Peroxidase 24	0.10	0.0017	0.0073	4.39	0.66
LGJ	1950000	lgj_g1644	Topless-related protein 2	0.10	0.0008	0.0017	2.06	0.88
LGJ	16440000	lgj_g2112	Transcription factor VRN1	0.23	0.0024	0.0043	1.76	0.52
LGJ	16890000	lgj_g2169	Universal stress protein PHOS34	0.17	0.0019	0.0030	1.63	0.39
LGJ	22590000	lgj_g2911	Major allergen Pru1	0.11	0.0006	0.0014	2.44	0.82
LGJ	47610000	lgj_g6191	Transcription factor ERF3	0.32	0.0012	0.0006	0.54	0.62

Table 5.6 continued

LG	Start	Gene name[a]	Inferred function[b]	het_cm[c]	pi-cm[d]	pi-cd[e]	cd/cm[f]	F_{ST}[g]
LGK	19140000	lgk_g2364	Receptor kinase-like protein Xa21	0.24	0.0023	0.0022	0.93	0.62
LGK	**25200000**	**lgk_g3168**	**Floral homeotic protein AGAMOUS**	**0.19**	**0.0010**	**0.0026**	**2.55**	**0.72**
LGK	26487000	lgk_g3344	GUN25_ARATH	0.23	0.0033	0.0059	1.76	0.39
LGL	18850000	lgl_g2401	protein kinase inhibitor SMR3	0.14	0.0012	0.0010	0.80	0.79
LGL	25960000	lgl_g3285	Multidrug resistance protein	0.17	0.0023	0.0007	0.31	0.79
LGL	37920000	lgl_g4777	Peroxidase 5	0.14	0.0031	0.0025	0.80	0.73
LGL	38090000	lgl_g4810	1-aminocyclopropane-1-carboxylate synthase 7	0.24	0.0059	0.0038	0.65	0.50
			Average over predicted genes within domestication regions	**0.13**	**0.0013**	**0.0025**	**1.88**	**0.72**
			Average over all predicted genes	**0.24**	**0.0031**	**0.0030**	**0.96**	**0.53**
			Standard deviation among all predicted genes	0.15	0.0038	0.0039	na	0.25

[a]AUGUSTUS predicted gene number; [b]Function inferred from Uniprot entry for the top protein alignment of the predicted gene; [c]proportion of heterozygous SNPS within predicted exons of the designated predicted gene for Chinese chestnut genome assemblies (n=17); [d]nucleotide diversity for Chinese chestnuts; [e]Nucleotide diversity within the predicted gene for American chestnuts (n=3); [f]Ratio of pi in American chestnut to Chinese chestnut; [g]*F*ST calculated between American and Chinese chestnuts.

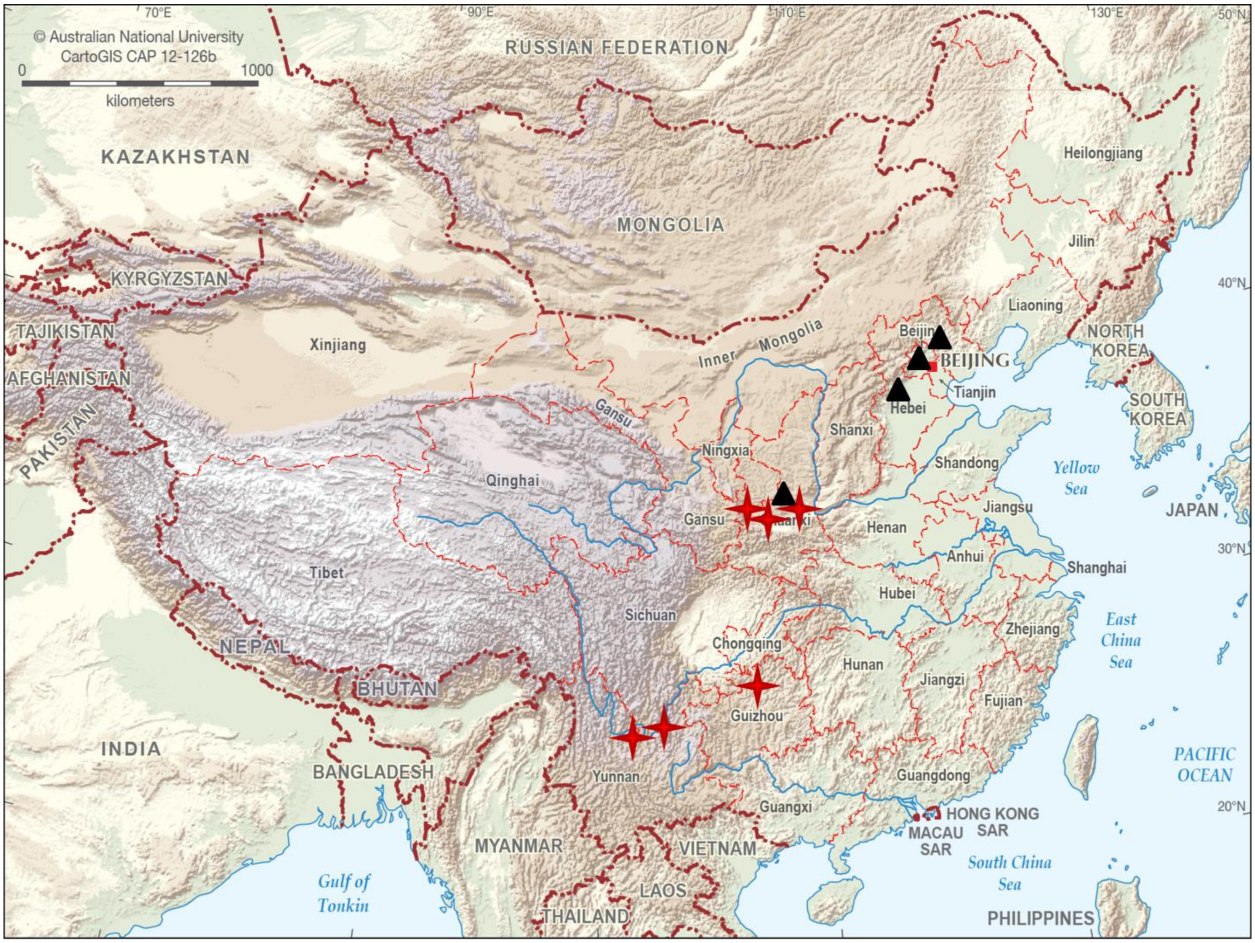


Figure 5.1 Map of the People's Republic of China showing locations from which wild trees were sampled (red stars) and the location of orchards sampled (black triangles).

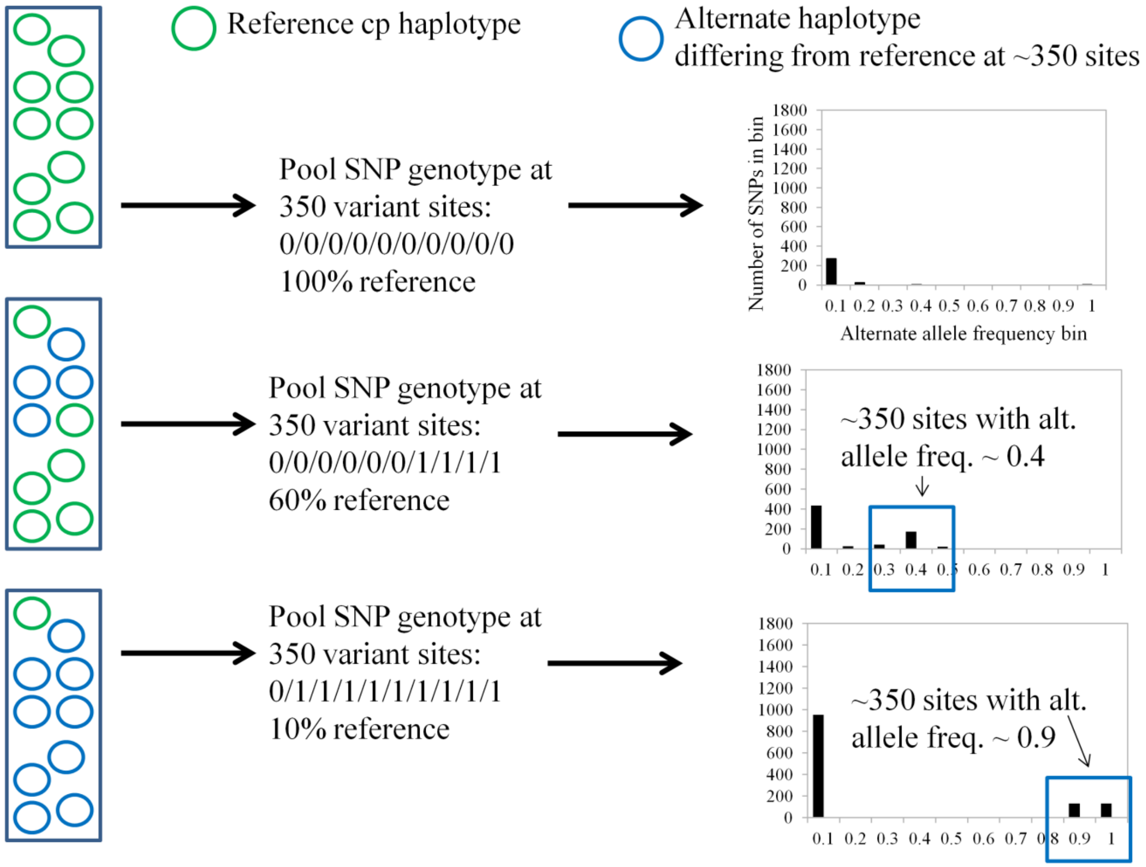

Reference cp haplotype
Alternate haplotype differing from reference at ~350 sites
Pool SNP genotype at 350 variant sites: 0/0/0/0/0/0/0/0/0/0 100% reference
Number of SNPs in bin
1800
1600
1400
1200
1000
800
600
400
200
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Alternate allele frequency bin
Pool SNP genotype at 350 variant sites: 0/0/0/0/0/0/1/1/1/1 60% reference
~350 sites with alt. allele freq. ~ 0.4
Pool SNP genotype at 350 variant sites: 0/1/1/1/1/1/1/1/1/1 10% reference
~350 sites with alt. allele freq. ~ 0.9

Figure 5.2 Depiction of alternate chloroplast haplotype frequency estimation from pooled samples. When SNP loci in the chloroplast are binned by the frequency of alternate alleles, the number of SNPs at which a given alternate haplotype differs from the reference (Y axis of histogram) and the frequency of the alternate genotype (X axis of histogram) can be estimated.

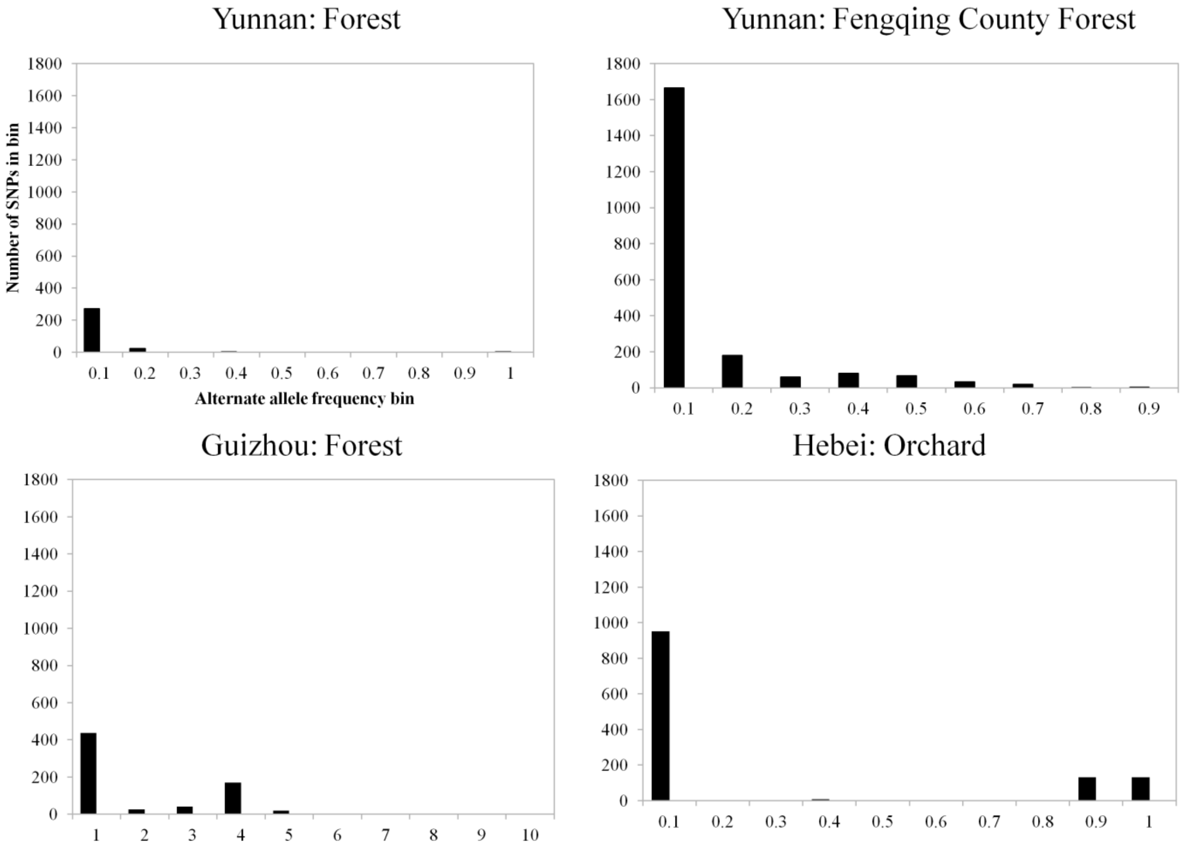


Figure 5.3 Chloroplast SNP alternate allele frequency histograms, with alternate allele frequency on the X axis and number of SNPS with a given alternate allele frequency on the Y axis, for pooled chestnut samples from southern (Yunnan, Guizhou) and northern (Hebei) China. An alternate chloroplast occurs at high frequency in the Hebei sample, while the reference chloroplast dominates one Yunnan sample and several haplotypes may be present in the Yunnan-Fengqing County sample.

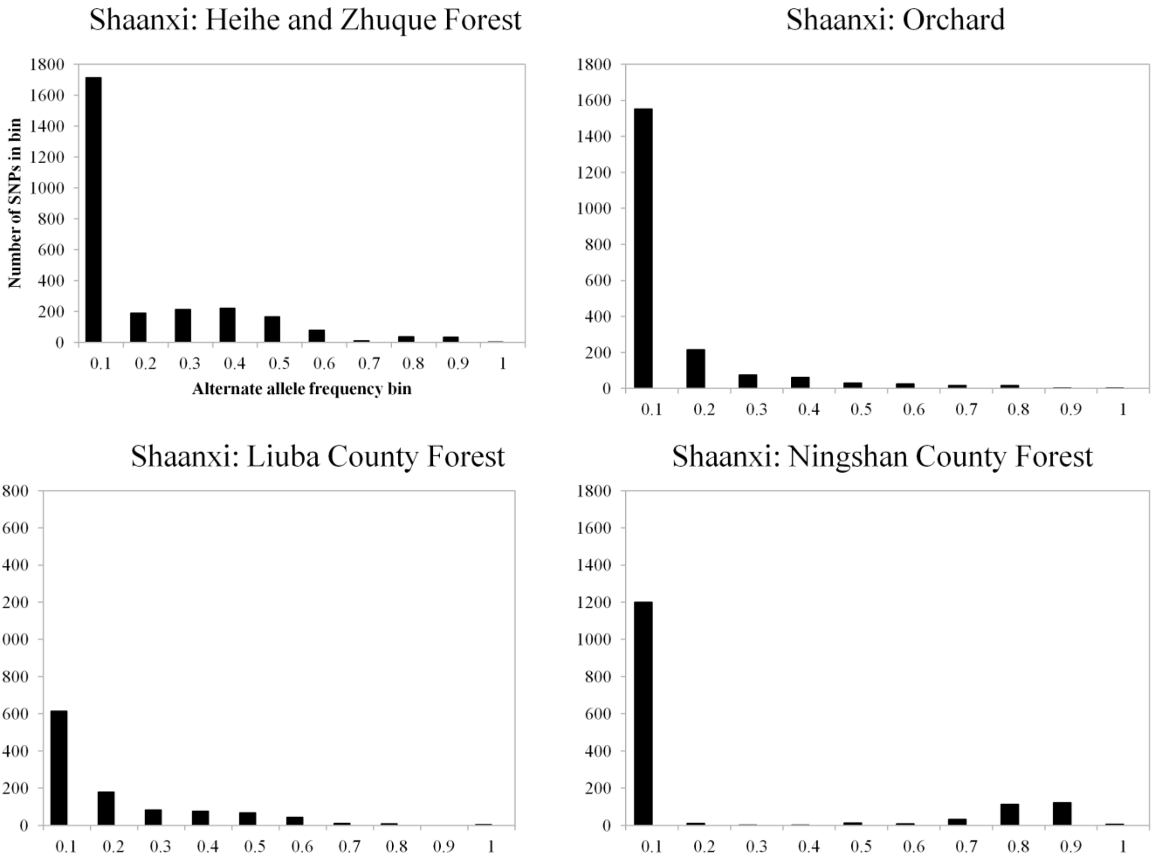


Figure 5.4 Chloroplast SNP alternate allele frequency histograms, with alternate allele frequency on the X axis and number of SNPS with a given alternate allele frequency on the Y axis, for pooled chestnut samples from Shaanxi Province in northwestern China. At least two haplotypes (peaks at frequency = 0.4 and = 0.8 are evident in the Heihe/Zhuque forest sample.