UNCOVERING GENES CONTROLLING APPLE TREE ARCHITECTURE AND FRUIT QUALITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Laura Elizabeth Dougherty

December 2019

UNCOVERING GENES CONTROLLING APPLE TREE ARCHITECTURE AND FRUIT QUALITY

Laura Elizabeth Dougherty, Ph. D.

Cornell University 2019

As the world population increases, there is a need to produce more food in limited land space. There is a strong need to optimize food production while reducing space, cost and food waste. This can be achieved with high density plantings, automated harvesting and food with longer shelf life. Apples are one of the most widely grown foods in the world and contain necessary nutrients. Apple trees have distinct, naturally occurring growth habits that include standard, weeping and columnar. Understanding the genetic mechanisms behind these growth habits can lead to more trees planted per acre and mechanized harvesting. In this dissertation, weeping (*W)* was mapped to chromosome 13 and *MdLazy1* was identified as a strong candidate gene underlying the weeping growth habit. Genetic mapping within a segregating columnar population revealed two recessive repressors regions of the columnar growth habit.

Apple fruit are stored in cold storage after harvest allowing them to be available year round for consumption. A major problem with fruit storage is the softening of fruit over time and a decrease in quality. In this dissertation, alleles of two ethylene synthesis genes *MdACS1* and *MdACS3a* were identified in 952 cultivars. Post-harvest storage studied on 131 cultivars combined with allelic genotypes revealed beneficial allelic combinations of the two genes that reduced fruit soften.

Research conducted and summarized in this dissertation addresses apple tree architecture and fruit quality. The genetic regions identified, makers analyzed, markers created and genes

characterized increase understanding of genetic mechanisms underlying weeping and columnar growth habit and fruit softening. This research helps address solutions to optimize food production, while also providing fruit breeder's information that can be utilized when creating new superior apple cultivars.

BIOGRAPHICAL SKETCH

Laura Elizabeth Dougherty was born and raised in Geneva, NY. She graduated from Geneva High School in 2007. She attended Keuka College (Keuka Park, NY) for her undergraduate education. She was a member of the women's lacrosse team and graduated in 2011 with degrees in biochemistry and business management. In the summer of 2010 between her junior and senior year she was a summer scholar in the plant pathology section at Cornell University, New York State Agricultural Experiment Station, Geneva, NY (renamed Cornell AgriTech in the summer of 2018). Her research experience there was eye opening and inspired her to chase a career in plant science. She was a laboratory technician from 2011-2013 in the horticulture section at Cornell University, New York State Agricultural Experiment Station, Geneva, NY. She attended Stony Brook University (Stony Brook, NY) from 2013-2014 completing her master's degree in biochemistry and cell biology. In 2015 she began her PhD studies at Cornell University (Ithaca, NY) in the School of Integrative Plant Science, horticulture section under Dr. Kenong Xu's direction. Her lab research was carried out at Cornell University, New York State Experimental Station, Geneva, NY.

ACKNOWLEDGMENTS

I firstly thank Dr. Kenong Xu for being my advisor and mentor. I am grateful for his guidance and support. I would also like to thank my committee members, Dr. Michael Scanlon and Dr. Haiyuan Yu for their insightful advice and guidance. I am grateful for all past and present members of the Xu lab for their support, assistance and most importantly friendships.

I would like to thank all my friends and family for their support. I am sorry for everything I missed because I was busy in lab. Your support never wavered and that means so much to me. Finally, I would like to thank my parents for all their support and encouragement throughout my life. You motivated me to be better and achieve my goals. Thank you for helping me move countless times and for always bringing me food.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Overview of Apple

Apples (genus *Malus)* are a member of the Rosaceae family, subfamily Pomoideae. They are a diverse group of over thirty species (Hancock, Luby, Brown, & Lobos, 2008). The United States Department of Agriculture (USDA) apple germplasm repository is home to 6,883 accessions from different *Malus* species around the world (Fazio, Forsline, Aldwinckle, & Pons, 2008). Domestic apple (*Malus domestica)* originated mainly from the progenitor species *M. sieversii* with intensive introgression from *M. sylvestris* (Duan et al., 2017). *M. sieversii* are wild apples native to Kazakhstan and central Asia (Volk, Henk, Richards, Forsline, & Chao, 2013). Genetically, *Malus* are primarily a highly heterozygous diploid organisms (2n=2x=34), although triploid and tetraploid cultivars also exist (S. K. Brown, 1992) . There is evidence that the apple genome underwent a genome wide duplication less than 50 million years ago, increasing from nine to eighteen chromosomes, then lost the eighteenth chromosome through translocation as homology is observed between chromosomes such as chromosomes 9 and 17 and chromosomes 4 and 12 (Velasco et al., 2010).

The apple genome is approximately 750Mb (million base pairs) in size encoding over 40,000 predicted genes. The first version of the apple reference genome from cultivar 'Golden Delicious' had 57,386 predicted genes in 742.3Mb (Velasco et al., 2010). In 2017 a 'Golden Delicious' double haploid was sequenced, predicting 42,140 genes in 651Mb (Daccord et al., 2017). Recently in 2019 a trihaploid 'Hanfu' apple with parentage 'Dongguang' x 'Fuji' was sequenced with 44,677 predicted genes in 708.54Mb (Zhang et al., 2019) as well as a *M. baccata* wild apple of 46,114 genes in 778Mb respectively (Chen et al., 2019; Zhang et al., 2019). As

sequencing assemblies and gene prediction algorithms continue to improve the total number of genes and genome sizes will continue to change. The function of over 10,000 identified genes remain unknown.

Apples as a fruit crop

Domestic apples are an important and versatile fruit crop. They are used for fresh consumption, and are processed into juices, ciders, vinegars and wines (Hancock et al., 2008). Apples are grown world-wide in temperate climates in over ninety countries. China is the largest producer in the world followed by the United Sates. In 2017 the United States produced 5,173,670 tons of apples (FAOSTAT, 2017). In 2016 the United States produced 5,160,750 tons of apples and exported 776,652 tons (FAOSTAT, 2017). The U.S. apple industry is estimated to be worth four billion dollars annually (USApple, 2019). In the United States there are over 200 apple cultivars grown, but the top produced apples are 'Gala', 'Red Delicious', 'Granny Smith', 'Fuji' and 'Honeycrisp', 'Golden Delicious', 'McIntosh', 'Rome', 'Cripps Pink/Pink Lady®', and 'Empire'(USApple, 2019).

Apple fruit quality and tree traits

Breeders are constantly striving to develop better quality apples to meet consumer needs. Apple fruit quality is a combination of taste, texture, shape and size of fruit (Hancock et al., 2008). Breeders must take into account regional preferences when creating new apple varieties. For example, in the United States a tarter apple is preferred while Asian markets favor sweet, low acid apples (Janick & Moore, 1996). Consumers perceive blemish free apples to be of high quality, and prefer round, large apples. Consumers initially assess fruit quality by appearance.

The USDA grades apples as U.S. Extra Fancy, U.S. Fancy, U.S. No. 1 or U.S. Utility based on appearance/size (USDA, 2019). The U.S Extra Fancy grade apples are the highest grade and fetch the highest sale prices for growers.

Apples are available year round, and must be stored properly to maintain quality during storage. Growers use controlled atmosphere storage units to store apples. In these controlled climates the oxygen and carbon dioxide levels are fine tuned to optimize storage quality and therefore apple shelf life (Siddiqui, Brackmann, Streif, & Bangerth, 1996). Apples are climacteric fruit that require ethylene to ripen, however in storage, the apples internal ethylene concentrations build up leading to fruit softening. Storage treatments with 1-methylcyclopropene (1- MCP), an ethylene inhibitor, can reduce the rate of softening, but not prevent it from occurring (Fan, Blankenship, & Mattheis, 1999).

Physiological disorders of apple fruit often appear during storage. Bitter pit, caused by calcium deficiency, is characterized by sunken lesions on apple skin and flesh below the skin can be dry/corky (Rosenberger, Schupp, Hoying, Cheng, & Watkins, 2004). Lenticel breakdown, is characterized by brown skin spots that enlarge overtime in storage as fruit soften. The exact cause of lenticel breakdown is unknown, although it is thought to be a combination of pre-harvest and post-harvest cultural practices (Turketti, Curry, & Lötze, 2012). Senescent breakdown, causes cortical browning of the apple flesh and skin. It is caused by mineral imbalance and calcium deficiency (Perring, 1968). Prolonged storage can leads to a greater severity of senescent breakdown. These disorders occur in major apple cultivars such as ‘Honeycrisp’, ‘Gala’, ‘Fuji’ and ‘Granny Smith’ and manifest on apple skin, reducing the quality of the fruit and therefore grade/price (USDA, 2019). The longer the fruit are in storage, the higher the chance of disorder development.

Disease resistance is a highly desired apple tree trait, as many diseases are devastating to fruit production (Hancock et al., 2008). The most devastating apple tree diseases are fire blight caused by *Erwinia amylovora*, apple scab caused by *Venturia inaequalis* and apple replant disease caused by a pathogen complex of fungi, oomyctes and nematodes (MacHardy, 1996; Mazzola, 1998; Vanneste, 2000). Fire blight and apple scab require pruning to remove diseased branches and in some cases full tree removal, reducing yields. Apple replant disease can cost growers up over $70,000 per acre during the first four years of production due to reduced productivity (Mazzola & Brown, 2010; Mazzola & Mullinix, 2005; Reed & Mazzola, 2015). Diseases that affect fruit appearance and quality include apple scab, black rot and sooty blotch/flyspeck (Venkatasubbaiah, Sutton, & Chilton, 1991; Williamson & Sutton, 2000).

Importance of Rootstocks

Apples are self-incompatible, meaning that they cannot self-pollinate (Broothaerts & Van Nerum, 2002). Seeds planted from an apple cultivar will not produce similar progeny. To produce more trees of a particular cultivar, vegetative propagation is commonly practiced. A cutting or bud from the desired variety is grafted onto a rootstock and grown. Grafting can be traced back over 2,000 years ago to the Romans (Webster, 1995). There are a large variety of rootstocks available that affect scion vigor, flowering, fruit set and yield while being disease resistance although the genetic mechanisms of how rootstocks influence the scions are unknown (Webster, 1995). The East Malling Research Station in England began rootstock research early in the 20th century (Tubbs, 1951). The Malling rootstock series offers size control, early fruiting and resistant to wooly apple aphids, however the rootstock is susceptible to fire blight (Extention, 2018). The Geneva rootstock series, developed in collaboration between Cornell

University and the United States Department of Agriculture-Agriculture Research Service (USDA-ARS), offer a range of tree vigor/height potential, resistance to fire blight, cold hardiness, high yield efficiency and tolerance to replant disease (Robinson, Aldwinckle, Fazio, & Holleran, 2002).

Apple production costs

Apple growing is a labor intensive process, making the cost of production high. Penn State extension estimates the cost of production for one acre of trees is $4,000-$5,000 per year for an established orchard. (Extension, 2017). Apple trees are typically pruned in the winter and occasionally in the summer to allow better sunlight penetration on fruit (Extension, 2017). Apple trees are supported by posts or trained on trellis systems, and in June fruit are thinned out to promote fruit size and prevent the trees from over bearing, ensuring a good fruit set the following year. Numerous sprays are applied throughout the growing season to prevent/control diseases, and all fruit are manually harvested in the fall. The amount of time and labor required for apple production is high. There is a strong push in the industry to automate labor intensive pruning and harvesting. It is estimated that pruning accounts for 20-25% of growers' production costs (Herrick, 2017). Prototypes for robotic pruners are being developed and tested but the unpredictability of tree shape is hard to program for (Herrick, 2017). During the 2019 harvest season in New Zealand, the first robotic harvester was used, however the orchard was planted to accommodate the harvester, with high density plantings and a 2D training system (Herrick, 2019).

Training systems

A limiting factor to tree productivity is light interception (Lakso & Robinson, 1996). Training systems involve the manipulation of branches to optimize fruit quality and tree productivity. Studies have been conducted on the efficiency of training systems over the years. There are two general canopy shapes, conic and V-shaped (Lordan, Gomez, Francescatto, & Robinson, 2019). Within the conic shapes there are slender pyramid, vertical axis, tall spindle and super spindle. In the V-shapes there are V-trellis, V-slender, V-tall spindle and V-super spindle (Lordan et al., 2019). Numerous studies have examined light interception, yield, and planting densities using different training systems, to determine which system is most productive (Hampson, Quamme, & Brownlee, 2002; Kappel & Quamme, 1993; Licznar-Małańczuk, 2004; Lordan et al., 2019). The results varied based on cultivar planted. Optimal profitability for 'Empire' apple was a conic shape planted at 2000 trees per hectare (2.47 acres), while 'Gala' was a conic shape planted with 3000 trees per hectare (Lordan et al., 2019).

Tree Architecture

Apple trees have different architectural shapes or growth habits, including standard, columnar, and weeping. Standard trees have many branches, mostly upright that expand vertically and horizontally creating a large full shaped canopy. Standard trees require extensive pruning to control tree shape and height. Columnar trees are compact, with thick stems, short internodes and limited upright branching, making them ideal for high density plantings. Their limited branching and exclusive fruiting on old wood spurs, require less pruning and upkeep than standard trees. Columnar trees arose from a somatic mutation of 'McIntosh' known as 'Wijcik McIntosh' (Wolters, Schouten, Velasco, Si-Ammour, & Baldi, 2013). Weeping trees have

downward growing branches and wide branching angles. Weeping growth habit is commonly found in crabapples, and desired in ornamental landscapes. Additionally, a genetic dwarf tree has very slow growth and short stature. Standard growth habit is most common, but in the right genetic backgrounds columnar and weeping growth habits are dominant, while the genetic dwarf is recessive (S. K. Brown, 1992).

Genetic mapping of traits

After a breeder makes an initial cross, it can take 20-25 years for one of the progeny to make it to the commercial market (Kellerhals & Meyer, 1994). Initial crosses are made between two parents and seeds are harvested. The minimum number of seeds suggested to harvest from a cross is 200-300, but the actual number of seeds harvested can be in the thousands (Hancock et al., 2008). Collected seeds then undergo vernalization before germination. After germination, evaluation for new varieties occurs in three steps. Step one is the initial planting of seedlings. Step two is propagation of promising seedlings by grafting. Step three is pre-commercial testing where more trees are propagated and grown at multiple test sites throughout the country, scaling up production (Hancock et al., 2008). Apples have a long juvenile phase that can last anywhere from 3 to 10 years before fruit set (Janick & Moore, 1996). The initial evaluation of seedlings in step one requires a considerable amount of resources, time and land. It is critical for breeders to efficiently make selections during step one to avoid carrying over undesirable seedlings to steps two and three. Understanding the genetics underlying important traits greatly helps breeders make informed selections. Genetic linkage maps have been generated for many fruit quality traits, tree architecture traits and disease resistance (Dolega, Dilworth, Koller, Gessler, & Kellerhals, 1999; Kenis & Keulemans, 2007; Liebhard, Kellerhals, Pfammatter, Jertmini, &

Gessler, 2003). Marker assisted breeding is using genetic genotypes to make informed selections. Markers tightly linked to traits of interests have been identified, helping breeders screen the seedlings quickly in step one, saving time, resources and land.

Apple acidity is a major component in apple taste and therefore quality. For fresh market consumption an acceptable titratable acidity (TA) range is 3-10mg/ml or pH 3.1-3.8 (A. G. Brown & Harvey, 1971; NYBOM, 1959; Visser & Verhaegh, 1978). Fruit with TA and pH readings outside of those ranges are unmarketable for either being too acidic or flat in taste. *Ma1*, a major gene in fruit acidity, was characterized, revealing high and low acid alleles (Bai et al., 2012). High acid genotypes (*Ma_*) and low acid genotypes (*mama*) can easily be identified with a cleaved amplified polymorphic sequences (CAPs) marker. The low acid genotypes fall below the acceptable TA range and therefore are undesired. This marker requires a small amount of DNA that can be taken from leaves weeks after planting, allows the undesirable *mama* seedlings to be discarded very early in step one, instead of waiting 3-10 years for the fruit to develop.

Almost every commercial variety of apple is susceptible to apple scab (M. L. Xu & Korban, 2000). Spray management for apple scab is costly and labor intensive with 12-15 sprays applied yearly (M. L. Xu & Korban, 2000). Six independent genes that confer resistance to apple scab have been identified (Williams & Kuc, 1969). Of those resistance genes, only the *Vf* gene (also known as *Rvi6*) was originally introgressed into susceptible varieties (Korban, 1998). In 1993 *Vf* gene resistance was overcome by a new race of *V. inaequalis* (Parisi, Lespinasse, Guillaumes, & Krüger, 1993). To increase resistance gene pyramiding strategies are now being applied to stack other scab resistant genes in addition to *Vf* (Baumgartner, Patocchi, Frey, Peil, & Kellerhals, 2015). Two Sequence Characterized Amplified Region (SCAR) markers have been designed that flank the *Vf* gene and are tightly linked to resistance. Use of the SCAR markers,

can determine if seedlings are homozygous resistant for the *Vf* gene (Tartarini, Gianfranceschi, Sansavini, & Gessler, 1999). Additional markers have been created for other apple scab resistance genes (Galli, Broggini, Kellerhals, Gessler, & Patocchi, 2010; Gessler, Patocchi, Sansavini, Tartarini, & Gianfranceschi, 2006). The use of resistance markers in progeny screening can increase the probability of new scab resistant cultivars by determining successful resistance gene stacking.

Marker assisted selection is a simple, cost effective way for breeders to make informed decisions about initial parental crosses and progeny selection when creating new varieties. As more genes of interest are identified and markers are made available, incorporating their use into breeding programs will lead to new superior varieties that will benefit growers and consumers.

Genome editing technology

Genome editing has been a major break though in the fields of genomics and biotechnology. In 2011 genome editing with nucleases (zinc finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and mega nucleases) was named Nature method of the year ("Method of the Year 2011," 2012). Genome editing encompasses targeted mutagenesis, gene deletions, gene disruption and gene addition (Tovkach, Zeevi, & Tzfira, 2009). In 2013 clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nucleases were first presented for genome editing (Cong et al., 2013). The editing methods have different modes of action, but all create double stranded DNA breaks (DSBs) at a targeted site, which triggers the cells to repair the DSBs with non-homologous end joining or homologous repair (Mushtaq et al., 2019). CRISPR/Cas9 is the most popular editing method, as multiple genes can be targeted at once (Bortesi & Fischer, 2015). In apple CRISPR/Cas9 was first used to

target phytoene desaturase (PDS) which is required for chlorophyll biosynthesis. Successfully edited plants were albino and easy to identify. This proof of concept study has opened the door for CRISPR/cas9 editing in apples (Nishitani et al., 2016).

While the potential applications of genome editing are exciting, one major concern with genome editing is the effects of foreign DNA integrated in the plant genome on human health/food safety. There is also concern of transgene introgression into wild non-transgenic crops (Stewart, Halfhill, & Warwick, 2003). Any plant with foreign DNA is classified as a genetically modified organism (GMO). The debate for/against GMOs is a hot button issue, and GMOs are under strict regulation (Wasmer, 2019; Yang & Chen, 2016). In the United States, three federal agencies regulate GMOs: USDA Animal and Plant Health Inspection Service (APHIS), the Food and Drug Administration (FDA), and the Environmental Protection Agency (EPA) (Kenong Xu, 2015). The Arctic apple developed by Okanagan Specialty fruits Inc., is the first transgenic apple approved for market in 2015 after going through a five year deregulation process through the USDA (Waltz, 2015). The novelty of the Arctic apple is that its flesh does not brown when cut, making it attractive for the prepared apple slices market. To eliminate browning, gene suppression through RNA interference was used to silence polyphenol oxidases, which interacts with phenols when apples are sliced and produce a compound of pigments primarily made up of quinones that gives the apples there unattractive brown appearance (K Xu, 2013). First available to the public in 2018, time will tell if the Arctic apple will be accepted by consumers.

Scientists are attempting to bypass GMO regulation, by using transient expression to edit plants, instead of genome integration. However studies have found DNA can still be integrated into the genome from transient vectors (Metje-Sprink, Menz, Modrzejewski, & Sprink, 2019).

DNA free genome editing, eliminating transgenes and plasmids, subsequently bypassing GMO regulations involves using proteins and RNA directly (Metje-Sprink et al., 2019). In 2016, apple protoplast were edited, targeting DIPM-1 to increase resistance to fire blight, using ribonucleoproteins (Malnoy et al., 2016). This is a promising step towards transgene free edited apples and increasing apple tree quality, however plant regeneration from protoplasts is difficult and brings challenges of its own.

Application of genome editing technologies in the future, can lead to development of new apple varieties with desirable traits faster than conventional breeding and allow for genetic improvement of established varieties. Genome editing can also be used to understand gene functions through reverse genetics. While the future of genome editing is promising, there must be a reduction in off-target editing, and additional research is needed to improve DNA free editing.

Future of apples

Apples are an important world commodity. Breeders continue to create new varieties that meet the needs of consumers and growers. New varieties like Rave™, SnapDragon®, Ruby Frost® and Cosmic Crisp® are creating waves of excitement and buzz for the apple industry (Jarvis, 2019; Palmer, 2013; Whitney, 2017). The hard cider and craft cider industries recently experienced a large resurgence as sales increased from $89.9 million in 2011, to $326.9 million in 2015 in the United States (Raboin, 2017). The booming cider market has opened up new research into cider/processing apples (Wattenberg, 2019; Weybright, 2018). Orchards are beginning to be planted and trained for automation as guidelines for mechanization are already available to growers (Alexander, Scheenstra, Miles, Musacchi, & King, 2019).

Researchers continue to study disease resistance, fruit texture, acidity, post-harvest

storage and other traits of interest. Understanding the mechanisms behind important traits will aid in development of markers for genetic selection, helping breeders make informative crosses and selections. Knowledge of which genes or specific alleles contribute to traits, will also be useful for targeted genome editing. In the following chapters, research was completed to further genetic understanding of apple tree architecture and post-harvest storage traits. In chapter 2, four genetic regions of interest for the weeping growth habit in *Malus* are identified including *W (Weeping)* on chromosome 13. In chapter 3, *MdLazy1* is identified within *W* as a strong candidate gene responsible for the weeping growth habit. Chapter 4 identifies two recessive genetic regions called *c2* and *c3* on chromosomes 10 and 9 respectively that can repress the columnar growth habit. In chapter 5, apple fruit were evaluated for fruit firmness and ethylene production 20 days post-harvest. Two new CAPs markers were created to distinguish between two null alleles of *MdACS3a.* The best allelic combination for reduced ethylene production, and limited fruit softening are *MdACS1-2* and *Mdacs3a.*

REFERENCES

Alexander, T. R., Scheenstra, E. J., Miles, C. A., Musacchi, S., & King, J. (2019). Establishing a cider apple orchard for mechanized management.

Bai, Y., Dougherty, L., Li, M., Fazio, G., Cheng, L., & Xu, K. (2012). A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the Ma locus is associated with low fruit acidity in apple. *Mol Genet Genomics, 287*.

Baumgartner, I. O., Patocchi, A., Frey, J. E., Peil, A., & Kellerhals, M. (2015). Breeding Elite Lines of Apple Carrying Pyramided Homozygous Resistance Genes Against Apple Scab and Resistance Against Powdery Mildew and Fire Blight. *Plant Molecular Biology Reporter, 33*(5), 1573-1583.

Bortesi, L., & Fischer, R. (2015). The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnology Advances, 33*(1), 41-52.

Broothaerts, W., & Van Nerum, I. (2002). *Apple self-incompatibility genotypes: an overview*. Paper presented at the XXVI International Horticultural Congress: Genetics and Breeding of Tree Fruits and Nuts 622.

Brown, A. G., & Harvey, D. M. (1971). Nature and inheritance of sweetness and acidity in cultivated apple. *Euphytica, 20*.

Brown, S. K. (1992). Genetics of Apple. In *Plant Breeding Reviews* (Vol. 9, pp. 333-366): John Wiley & Sons, Inc.

Chen, X., Li, S., Zhang, D., Han, M., Jin, X., Zhao, C., Wang, S., Xing, L., Ma, J., Ji, J., & An, N. (2019). Sequencing of a Wild Apple (Malus baccata) Genome Unravels the Differences Between Cultivated and Wild Apple Species Regarding Disease Resistance and Cold Tolerance. *G3: Genes|Genomes|Genetics, 9*(7), 2051-2060.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science, 339*(6121), 819-823.

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., & Velasco, R. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*.

Dolega, E., Dilworth, E., Koller, B., Gessler, C., & Kellerhals, M. (1999). *ADVANCES IN MARKER-ASSISTED APPLE BREEDING*. Paper presented at the Eucarpia symposium on Fruit Breeding and Genetics 538.

Duan, N., Bai, Y., Sun, H., Wang, N., Ma, Y., Li, M., Wang, X., Jiao, C., Legall, N., Mao, L., Wan, S., Wang, K., He, T., Feng, S., Zhang, Z., Mao, Z., Shen, X., Chen, X., Jiang, Y., Wu, S., Yin, C., Ge, S., Yang, L., Jiang, S., Xu, H., Liu, J., Wang, D., Qu, C., Wang, Y., Zuo, W., Xiang, L., Liu, C., Zhang, D., Gao, Y., Xu, Y., Xu, K., Chao, T., Fazio, G., Shu, H., Zhong, G.-Y., Cheng, L., Fei, Z., & Chen, X. (2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nature Communications, 8*(1), 249.

Extension, P. (2017, 9-28-2017). Apple Production. *Agricultural Alternatives.* Retrieved from https://extension.psu.edu/apple-production

Extention, P. (2018). Apple Rootstocks: Capabilities and Limitations. Retrieved from https://extension.psu.edu/apple-rootstocks-capabilities-and-limitations

Fan, X., Blankenship, S. M., & Mattheis, J. P. (1999). 1-Methylcyclopropene Inhibits Apple Ripening. *124*(6), 690.

FAOSTAT. (2017). Food and agricultural organization of the United Nations. Retrieved from http://www.fao.org/faostat/en/

Fazio, G., Forsline, P., Aldwinckle, H. s., & Pons, L. (2008). The Apple Collection in Geneva, NY: A Resource for The Apple Industry Today and for Generations to Come.

Galli, P., Broggini, G. A. L., Kellerhals, M., Gessler, C., & Patocchi, A. (2010). High-resolution genetic map of the Rvi15 (Vr2) apple scab resistance locus. *Molecular Breeding, 26*(4), 561-572.

Gessler, C., Patocchi, A., Sansavini, S., Tartarini, S., & Gianfranceschi, L. (2006). Venturia inaequalis Resistance in Apple. *Critical Reviews in Plant Sciences, 25*(6), 473-503.

Hampson, C. R., Quamme, H. A., & Brownlee, R. T. (2002). Canopy Growth, Yield, and Fruit Quality of 'Royal Gala' Apple Trees Grown for Eight Years in Five Tree Training Systems. *37*(4), 627.

Hancock, J. F., Luby, J. J., Brown, S. K., & Lobos, G. A. (2008). Apples. In J. F. Hancock (Ed.), *Temperate Fruit Crop Breeding: Germplasm to Genomics* (pp. 1-38). Dordrecht: Springer Netherlands.

Herrick, C. (2017). Pruning Goes High-Tech. *Growing Produce*. Retrieved from https://www.growingproduce.com/fruits/apples-pears/pruning-goes-high-tech/

Herrick, C. (2019). Robotic Apple Harvester Makes Debut in New Zealand. Retrieved from https://www.growingproduce.com/fruits/apples-pears/robotic-apple-harvester-makes-debut-in-new-zealand/

Janick, J., & Moore, J. N. (1996). *Fruit breeding, tree and tropical fruits* (Vol. 1): John Wiley & Sons.

Jarvis, B. (2019). Cosmic Crisp Apple Launch. Retrieved from https://story.californiasunday.com/cosmic-crisp-apple-launch

Kappel, F., & Quamme, H. A. (1993). Orchard training systems influence early canopy development and light microclimate within apple tree canopies. *Canadian Journal of Plant Science, 73*(1), 237-248.

Kellerhals, M., & Meyer, M. (1994). Aims of the apple breeding programme at Wädenswil. In *Progress in Temperate Fruit Breeding* (pp. 117-121): Springer.

Kenis, K., & Keulemans, J. (2007). Study of tree architecture of apple (Malus× domestica Borkh.) by QTL analysis of growth traits. *Molecular Breeding, 19*(3), 193-208.

Korban, S. (1998). *What's new with disease-resistant apple cultivars.* Paper presented at the Proc. Trans. Ill. Hortic. Soc.

Lakso, A., & Robinson, T. (1996). *Principles of orchard systems management optimizing supply, demand and partitioning in apple trees.* Paper presented at the VI International Symposium on Integrated Canopy, Rootstock, Environmental Physiology in Orchard Systems 451.

Licznar-Małańczuk, M. (2004). Influence of planting and training systems on fruit yield in apple orchard. *J. Fruit Ornam. Plant Res, 12*, 97-104.

Liebhard, R., Kellerhals, M., Pfammatter, W., Jertmini, M., & Gessler, C. (2003). Mapping quantitative physiological traits in apple (Malus × domestica Borkh.). *Plant Molecular Biology, 52*(3), 511-526.

Lordan, J., Gomez, M., Francescatto, P., & Robinson, T. L. (2019). Long-term effects of tree density and tree shape on apple orchard performance, a 20 year study – part 2, economic analysis. *Scientia Horticulturae, 244*, 435-444.

MacHardy, W. E. (1996). *Apple scab: biology, epidemiology, and management*: American Phytopathological Society (APS Press).

Malnoy, M., Viola, R., Jung, M.-H., Koo, O.-J., Kim, S., Kim, J.-S., Velasco, R., & Nagamangala Kanchiswamy, C. (2016). DNA-Free Genetically Edited Grapevine and Apple Protoplast Using CRISPR/Cas9 Ribonucleoproteins. *Frontiers in Plant Science, 7*(1904).

Mazzola, M. (1998). Elucidation of the Microbial Complex Having a Causal Role in the Development of Apple Replant Disease in Washington. *Phytopathology, 88*(9), 930-938.

Mazzola, M., & Brown, J. (2010). Efficacy of brassicaceous seed meal formulations for the control of apple replant disease in conventional and organic production systems. *Plant Disease, 94*(7), 835-842.

Mazzola, M., & Mullinix, K. (2005). Comparative field efficacy of management strategies containing Brassica napus seed meal or green manure for the control of apple replant disease. *Plant Disease, 89*(11), 1207-1213.

Method of the Year 2011. (2012). *Nature Methods, 9*(1), 1-1.

Metje-Sprink, J., Menz, J., Modrzejewski, D., & Sprink, T. (2019). DNA-Free Genome Editing: Past, Present and Future. *Frontiers in Plant Science, 9*(1957).

Mushtaq, M., Sakina, A., Wani, S. H., Shikari, A. B., Tripathi, P., Zaid, A., Galla, A., Abdelrahman, M., Sharma, M., Singh, A. K., & Salgotra, R. K. (2019). Harnessing Genome Editing Techniques to Engineer Disease Resistance in Plants. *Frontiers in Plant Science, 10*(550).

Nishitani, C., Hirai, N., Komori, S., Wada, M., Okada, K., Osakabe, K., Yamamoto, T., & Osakabe, Y. (2016). Efficient Genome Editing in Apple Using a CRISPR/Cas9 system. *Scientific Reports, 6*, 31481.

NYBOM, N. (1959). On the inheritance of acidity in cultivated apples. *Hereditas, 45*(2-3), 332-350.

Palmer, R. (2013). New Apple Varieties We're Excited About: SnapDragon, Ruby Frost. *Internatinal Business Times.* Retrieved from https://www.ibtimes.com/new-apple-varieties-were-excited-about-snapdragon-rubyfrost-1436328

Parisi, L., Lespinasse, Y., Guillaumes, J., & Krüger, J. (1993). A new race of Venturia inaequalis virulent to apples with resistance due to the Vf gene. *Phytopathology, 83*(5), 533-537.

Perring, M. A. (1968). Mineral composition of apples. VII.—The relationship between fruit composition and some storage disorders. *Journal of the Science of Food and Agriculture, 19*(4), 186-192.

Raboin, M. (2017). Hard Cider in the North Central Region: Industry Survey Finsings. Retrieved from https://www.cias.wisc.edu/wp-content/uploads/2017/07/cideerstudy071817web.pdf

Reed, A., & Mazzola, M. (2015). Characterization of apple replant disease-associated microbial communities over multiple growth periods using next-generation sequencing. *Phytopathology, 105*, 117.

Robinson, T., Aldwinckle, H., Fazio, G., & Holleran, T. (2002). *The Geneva series of apple rootstocks from Cornell: performance, disease resistance, and commercialization.* Paper presented at the XXVI International Horticultural Congress: Genetics and Breeding of Tree Fruits and Nuts 622.

Rosenberger, D. A., Schupp, J. R., Hoying, S. A., Cheng, L., & Watkins, C. B. (2004). Controlling Bitter Pit in `Honeycrisp' Apples. *14*(3), 342.

Siddiqui, S., Brackmann, A., Streif, J., & Bangerth, F. (1996). Controlled atmosphere storage of apples: Cell wall composition and fruit softening. *Journal of Horticultural Science, 71*(4), 613-620.

Stewart, C. N., Halfhill, M. D., & Warwick, S. I. (2003). Transgene introgression from genetically modified crops to their wild relatives. *Nature Reviews Genetics, 4*(10), 806-817.

Tartarini, S., Gianfranceschi, L., Sansavini, S., & Gessler, C. (1999). Development of reliable PCR markers for the selection of the Vf gene conferring scab resistance in apple. *Plant Breeding, 118*(2), 183-186.

Tovkach, A., Zeevi, V., & Tzfira, T. (2009). A toolbox and procedural notes for characterizing novel zinc finger nucleases for genome editing in plant cells. *The Plant Journal, 57*(4), 747-757.

Tubbs, F. R. (1951). East Malling Research Station. *Proceedings of the Royal Society of London. Series B - Biological Sciences, 139*(894), 1-18.

Turketti, S. S., Curry, E., & Lötze, E. (2012). Role of lenticel morphology, frequency and density on incidence of lenticel breakdown in 'Gala' apples. *Scientia Horticulturae, 138*, 90-95.

USApple. (2019). Apple Industry at a Glance. Retrieved from http://usapple.org/the-industry/apple-industry-at-a-glance/

USDA. (2019). Apple Grades & Standards. Retrieved from https://www.ams.usda.gov/grades-standards/apple-grades-standards

Vanneste, J. L. (2000). *Fire blight: the disease and its causative agent, Erwinia amylovora*: CABI.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., & Kalyanaraman, A. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). *Nat Genet, 42*.

Venkatasubbaiah, P., Sutton, T., & Chilton, W. (1991). Effect of phytotoxins produced by Botryosphaeria obtusa, the cause of black rot of apple fruit and frogeye leaf spot. *Phytopathology, 81*(3), 243-247.

Visser, T., & Verhaegh, J. J. (1978). Inheritance and selection of some fruit characters of apple. I. Inheritance of low and high acidity. *Euphytica, 27*.

Volk, G. M., Henk, A. D., Richards, C. M., Forsline, P. L., & Chao, C. T. (2013). Malus sieversii: A diverse Central Asian apple species in the USDA-ARS national plant germplasm system. *HortScience, 48*(12), 1440-1444.

Waltz, E. (2015). Nonbrowning GM apple cleared for market. *Nature biotechnology, 33*(4), 326-327.

Wasmer, M. (2019). Roads Forward for European GMO Policy-Uncertainties in Wake of ECJ Judgment Have to be Mitigated by Regulatory Reform. *Frontiers in bioengineering and biotechnology, 7*, 132-132.

Wattenberg, I. (2019). Univ. of Minn. awarded $100k grant to research apples for state's cidermakers. *The Growler*. Retrieved from https://growlermag.com/univ-of-minn-awarded-100k-grant-to-research-apples-for-states-cidermakers/

Webster, A. D. (1995). Rootstock and interstock effects on deciduous fruit tree vigour, precocity, and yield productivity. *New Zealand Journal of Crop and Horticultural Science, 23*(4), 373-382.

Weybright, S. (2018). Grant supports improving cider industry-'Apple to Glass'. *WSU Insider.* Retrieved from https://news.wsu.edu/2018/10/12/grant-supports-improving-cider-industry/

Whitney, A. (2017). Rave Apples Are a Brand New Apple in Your Grocery Store. Retrieved from https://www.bonappetit.com/story/rave-apples

Williams, E., & Kuc, J. (1969). Resistance in Malus to Venturia inaequalis. *Annual Review of Phytopathology*, *7*(1), 223-246.

Williamson, S. M., & Sutton, T. B. (2000). Sooty blotch and flyspeck of apple: etiology, biology, and control. *Plant Disease*, *84*(7), 714-724.

Wolters, P. J., Schouten, H. J., Velasco, R., Si-Ammour, A., & Baldi, P. (2013). Evidence for regulation of columnar habit in apple by a putative 2OG-Fe(II) oxygenase. *New Phytol*, *200*(4), 993-999.

Xu, K. (2013). An overview of Arctic apples: Basic facts and characteristics. *New York Fruit Quarterly. New York State Horticultural Society, Fall*.

Xu, K. (2015). Arctic Apples: A Look Back and Forward. *Fruit Quarterly, 23(2)*.

Xu, M. L., & Korban, S. S. (2000). Saturation mapping of the apple scab resistance gene Vf using AFLP markers. *Theoretical and Applied Genetics*, *101*(5), 844-851.

Yang, Y. T., & Chen, B. (2016). Governing GMOs in the USA: science, law and public health. *Journal of the Science of Food and Agriculture*, *96*(6), 1851-1855.

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., Zhang, C., Tian, Y., Liu, G., Gul, H., Wang, D., Tian, Y., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., Wang, D., & Cong, P. (2019). A high-quality apple genome assembly reveals the

association of a retrotransposon and red fruit colour. *Nature Communications*, *10*(1), 1494.

CHAPTER 2

Exploring DNA variant segregation types in pooled genome sequencing enables effective mapping of weeping trait in *Malus*

Abstract

To unlock the power of next generation sequencing-based bulked segregant analysis in allele discovery in out-crossing woody species, and to understand the genetic control of weeping trait, an F_1 population from cross 'Cheal's Weeping' × 'Evereste' was used to create two genomic DNA pools 'weeping (17 progeny)' and 'standard (16 progeny)'. Illumina pair-end (2x151bp) sequencing of the pools to a 27.1x (weeping) and a 30.4x (standard) genome (742.3Mb) coverage allowed detection of 84,562 DNA variants specific to 'weeping', 92,148 specific to 'standard' and 173,169 common to both pools. A detailed analysis of the DNA variant genotypes in the pools predicted three informative segregation types of variants: <**l**m x mm> (Type-I) in weeping pool specific variants, and <**l**m x **ll**> (Type-II) and <**h**k x **h**k> (Type-III) in variants common to both pools. Note: the first allele is assumed weeping linked and the allele in bold is a variant to the reference genome. Conducting variant allele frequency and density-based mappings revealed four genomic regions with a significant association with weeping: a major locus *Weeping* (*W*) on chromosome 13 and others on chromosomes 10 (*W2*), 16 (*W3*) and 5 (*W4*). The results from Type-I variants were noisier and less certain than those from Type-II and Type-III variants, demonstrating that although Type-I variants are often the first choice, Type-II and Type-III variants represent an important source of DNA variants that can be exploited for genetic mapping in out-crossing woody species. Confirmation of the mapping of *W* and *W2*, investigation into their genetic interactions, and identification of expressed genes in the *W* and *W2* regions provided insight into the genetic control of weeping and its expressivity in *Malus*.

Introduction

The development of next generation sequencing (NGS) technologies has revolutionized approaches in genetics and genomics studies (Mardis, 2008; Quail et al., 2012; Schuster, 2008). Implementing the NGS technology enabled whole genome sequencing in bulk segregant analysis (BSA), a methodology (Giovannoni, Wing, Ganal, & Tanksley, 1991; Michelmore, Paran, & Kesseli, 1991) widely used in genetic mapping by analyzing two pools of genomes of contrasting phenotypes (Lister, Gregory, & Ecker, 2009; K. Schneeberger et al., 2009). This allows for gene/QTL mapping as constrasting DNA pools should have different underlying genetics in the locus of interest. Single nucledotide polymorphisms (SNPs) between pools can act as markers. Regions in close proximity to the causal mutation are homologus in one pool compared to the other making identification of the causal mutation easier as all the sequencing data is present. After the first successful demonstrations in *Arabidopsis thaliana* (Lister et al., 2009; K. Schneeberger et al., 2009), the approach has been adapted in many other species, such as legume (Sandal et al., 2012), rice (Abe et al., 2012), wheat (Trick et al., 2012), arthropods (Van Leeuwen et al., 2012), zebrafish (Obholzer et al., 2012) and peach (Dardick et al., 2013). Similar analyses using RNA-seq data were reported in maize and zebrafish (Hill et al., 2013; Liu, Yeh, Tang, Nettleton, & Schnable, 2012; Miller, Obholzer, Shah, Megason, & Moens, 2013). A few of the latest examples of using pooled genome sequencing analyses to identify important genes in plants include: the lettuce thermotolerant seed germination gene *ABA1/ZEP* (Huo et al., 2016), the glycerol-3-phosphate acyltransferase gene *GPAT6* crucial in tomato fruit cutin biosynthesis (Petit et al., 2016), and the gibberellic acid receptor *PpeGID1c* gene for brachytic dwarfism in peach (Hollender, Hadiarto, Srinivasan, Scorza, & Dardick, 2016). In addition, the approach has

been extended to mapping genetic variants associated with DNA methylation (Kaplow et al., 2015) and genome-wide association studies (Yang et al., 2015).

Varying terms have been used to describe the application of NGS-enabled whole genome sequencing in BSA, such as mapping-by-sequencing (Hartwig, James, Konrad, Schneeberger, & Turck, 2012), whole genome sequencing (Leshchiner et al., 2012; Sarin, Prabhu, O'Meara, Pe'er, & Hobert, 2008; Korbinian Schneeberger, 2014), pool-seq (Kofler, Pandey, & Schlötterer, 2011), MutMap (Abe et al., 2012), QTL-seq (Takagi et al., 2013), pnome (Dardick et al., 2013), and others. Regardless of terminology, the basic ideas and principles behind the pooled genome sequencing approach are similar, i.e. the genome pool from individuals with a trait of interest would have more abundant DNA molecules carrying the causal variants than the genome pool from those without the trait. As a result, the frequency of the causal variant or the linked variants is expected to be different from that in unlinked regions. In the case of a dominant trait in BC1 population, the causal variant frequency is expected to be ca. 50% in the pool with the trait, whereas the frequency in the pool without the trait will be ca. zero. The frequency of DNA variants towards both directions from the causal variant will progressively become lower than 50%, i.e. a causal mutation is most likely under the peak of DNA variant frequency in the pool carrying the trait of interest in this example.

To facilitate data analysis of pooled genome sequencing, several analytical software packages have been developed, such as SHOREmap which can identify SNPs and 1-3 basepair indels (insertions/deletions) that indicate frameshift mutations with high sensitivity (K. Schneeberger et al., 2009; Sun & Schneeberger, 2015).CloudMap , which can pinpoint sequence variations and generates a list of candidate variants for rapid identification (Minevich, Park, Blankenberg, Poole, & Hobert, 2012). SNPtrack, which is used for SNP discovery, mutation

localization and is capable of inferring recombinants, reducing the size of regions of interest. (Leshchiner et al., 2012). MegaMapper can facilitate haplotype calling, homozygosity measurements and filter out known wild type variants from screening (Obholzer et al., 2012). MMAPPR (Mutation Mapping Analysis Pipeline for Pooled RNA-seq) uses RNA sequencing data to determine allele frequencies, identification of mutation region and provides a list of putative mutations in the coding sequences (Hill et al., 2013). EXPLoRA (Extraction of overrepresented alleles) uses linkage disequalibrum through hidden markov model to identify QTLs and EXPLoRA-Web is a user friendly platform to run the algorithm (Duitama et al., 2014; Pulido-Tamayo, Duitama, & Marchal, 2016).Finally, GIPS (Gene Identification via Phenotype Sequencing) identifies SNPs and the significance of each candidate gene with the associated phenotype (Hu et al., 2016). To run all these analytical software packages, besides EXPLoRA-web, knowledge of programmer coding, such as R, C++ or Python is required. Each method differs in its approach to analyze variants, however these approaches all try to identify regions of interest and causal mutations.

A method for identification of genomic regions carrying a causal mutation in unordered genomes has also been developed for species without a reference genome (Corredor-Moreno, Chalstrey, Lugo, & MacLean, 2015). These packages are helpful tools for pooled genome sequencing data analysis for many model species from which they were developed. However, efforts are needed to make them more user-friendly and/or to broaden their application range to cover non-model species or species without a high-quality reference genome. In addition, accurate calling of variants remains challenging as a considerable fraction of variants that are false appeared to be inherent to commonly used variant callers (Huang, Mullikin, & Hansen, 2015; Ribeiro et al., 2015).

A number of mapping strategies for positioning causal variants using pooled genome sequencing data have been proposed and demonstrated with successful applications, such as variant scarcity or density mappings (K. Schneeberger et al., 2009; Zuryn, Le Gras, Jamet, & Jarriault, 2010), variant discovery mapping (Minevich et al., 2012), SNP index (Abe et al., 2012), bulk segregant linkage mapping (Obholzer et al., 2012), delta SNP index mapping (Fekih et al., 2013; Takagi et al., 2013), SNP ratio mapping (SRM) (Lindner et al., 2012), mutant allele frequency (MAF), allelic distance (AD), and homozygosity mapping (Korbinian Schneeberger, 2014). These mapping strategies largely can be attributed to the use of three major parameters, including variant allele frequency, variant density, and variant distance (allelic distance). It should be possible in principle to conduct pooled genome sequencing based genetic mapping studies in *Malus* species although the DNA variants are of complex segregation patterns and the phase is often unknown due to their heterogeneously heterozygous genome.

Weeping growth habit in woody species represents a unique form of tree architecture and has been an essential element in landscape aesthetics. Compared with standard trees with branches that grow mostly upward with certain angles, weeping tree branches grow downward. In *Malus*, weeping (pendulous) phenotype exists in *M. domestica,* such as cv 'Elisa Ratkee'. But it is more frequently seen in crabapples for ornamental purpose, such as 'Exzellenz Thiel', 'Red Jade', 'baccata 'Gracilis', 'Cheal's Weeping', and 'Louisa'. 'Red Jade' is believed to be derived from an open pollinated seedlings of 'Exzellenz Thiel', which was selected from cross *M. prunifolia* 'Pendula' x *M. floribunda* (S.K. Brown, Maloney, Hemmat, & Aldwinckle, 2004). The weeping phenotype in *M. baccata* 'Gracilis' is controlled by a single dominant allele, called *Weeping* (*W*), based on an inheritance study conducted in two small populations of 28 seedlings derived from *M. baccata* 'Gracilis' (Alston, Phillips, & Evans, 2000; Susan K. Brown, 1992;

Sampson & Cambron, 1965). In a population of 98 seedlings from cross ‘Wijcik McIntosh’ (columnar) × ‘Red Jade’ (weeping), the weeping and columnar phenotypes segregated independently despite intermediates expressing both phenotypes, i.e. columnar at top while weeping at bottom (Just, 2001). Based on the phenotype of weeping trees, the gene underlying *W* must be involved in positive shoot gravitropism and branch angle formation.

Cultivated apple tree architecture has been categorized into four types based on the overall growth habit: columnar (e.g. ‘Wijcik McIntosh’), spur (‘Starkrimson’), standard (‘Golden Delicious’) and weeping (‘Granny Smith’) (Costes, Lauri, & Regnard, 2006; Höfer et al., 2013; J. M. Lespinasse & Delort, 1986; Y. Lespinasse, 1992; Pereira-Lorenzo, Ramos-Cabrer, & Fischer, 2009). However, ‘Granny Smith’ trees grow branches similar to a standard tree and their classification as weeping is due to the bending of branches that bear fruit at their tips. When no fruit is present, the weeping phenotype cannot be observed. This is distinctly different from the weeping trait studied in this report where branches actively grow downward and are not a result of fruit weight. A better understanding of the genetic architecture responsible for the weeping trait in *Malus* would provide important insight into directional growth of shoot meristems in woody species.

In this study, based on a detailed analysis of DNA variant genotypes in the weeping and standard pools and their possible segregation types, an effective strategy was devised to target three informative segregation types of variants: <**l**m x mm> (Type-I) in weeping pool specific variants, and <**l**m x **ll**> (Type-II) and <**h**k x **h**k> (Type-III) in variants common to both pools. Note that the first allele is designated weeping linked from ‘Cheal’s Weeping’ and the alleles in bold represent a DNA variant in relation to the apple reference genome. Although Type-I variants are the most straightforward for mapping because they are exclusive to the weeping pool

and contain the causal mutation, Types II and III variants performed better in mapping the weeping trait identifying fewer genomic regions of interest, highlighting their utility in pooled genome sequencing-based genetic mapping. To the best of our knowledge, this is the first report identifying and exploiting three informative segregation types of DNA variants in pooled genome sequencing analysis for genetic mapping in an out-crossing woody species of highly heterozygous genome.

Materials and methods

Plant materials and growth habit evaluation

Three F_1 populations segregating for weeping growth habit were used for genetic mapping of the trait. The first comprised 38 seedling trees (8-year old) from cross 'Cheal's Weeping' × 'Evereste' (**Figure 2.1 i-ii**); the second was developed from NY-051 × 'Louisa' of 140 progeny; and the third was derived from NY-011 × NY-100 consisting of 39 individuals. The progeny in the second and third populations were 2-years old. 'Cheal's Weeping' and 'Louisa' are weeping crabapple cultivars. NY-100 is a weeping selection from the progeny of 'Red Jade', another weeping crabapple cultivar. The relatedness of 'Cheal's Weeping', 'Louisa' and 'Red Jade' is unknown. 'Everest', NY-051 and NY-011 are crabapples of standard growth habit. The populations were planted in a research orchard of Cornell University in Geneva, New York, USA. Evaluation of growth habits was conducted by visual observation and seedling trees were categorized into weeping, weeping-like, standard, standard-like and intermediate (**Figure S2.1**).

Construction and sequencing of genomic DNA pools

Genomic DNA samples were prepared from young leaf tissues as previously described (Wang et al., 2012) and were quantified using Qubit dsDNA BR Assay Kit on a Qubit 3.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). An equal amount of DNA (300 ng) from each of the 17 weeping (-like) and 16 standard progeny in population 'Cheal's Weeping' × 'Everest' was combined into a weeping pool and a standard pool, respectively (**Figure 2.1 ii**). Genomic DNA libraries of target insert size of 500 bp were constructed from each of the two genomic DNA pools using Illumina (San Diego, CA, USA) TruSeq DNA PCR-Free Library Preparation Kit, and then paired-end (2 x 151 bp) sequenced on an Illumina HiSEQ 2500 platform (**Figure 2.1 iii**) at the Genomics Facility of Cornell University (Ithaca, New York, USA).

Mapping of reads to the apple reference genome

The assembled apple 'Golden Delicious' genome MalDom1.0 (NCBI accession GCA_000148765.1, annotation release 100, June 2014) (Velasco et al., 2010), which comprises 17 chromosomes with a total size of 526,197,889 bp, was used as reference. Mapping of the Illumina sequencing reads onto the reference genome was conducted in the weeping and standard pools, respectively, using software CLC Genomics Workbench (v7.5, CLCBio, Cambridge, MA, USA). The mapping parameters and settings were similar to previously described (Bai, Dougherty, & Xu, 2014), i.e. the minimum length fraction is 0.8 and the minimum similarity is 0.98 (**Figure 2.1 iv**, **Table S2.1**).

Detection and analysis of DNA variants

In the weeping or standard pool, detection of DNA variants was conducted using the fixed ploidy (2x) variant detection tool embedded in CLC Genomics Workbench (**Figure 2.1 v**). Variant frequency was calculated automatically based on the total number of reads aligned at the region. To capture as many variants as possible initially, the minimum coverage was ten and the minimum count of variant reads was two. The variants were filtered through a series of filters to remove variants that are reference alleles, hyper allelic, homopolymers, and/or called when reference is an ambiguous base, such as M, R, W, S, Y and K (**Table S2.2**). Variants specific to either pool and variants common to both pools were identified by direct comparison between the set of variants identified in the weeping pool and those in the standard pool using CLC Genomic Workbench (**Figure 2.1 vi**). To minimize false positive variants prior to mapping the trait, these pool-specific and common variants were filtered again by another set of filters: read coverage ≥20; forward/reverse reads balance 0.25-0.5, number of reads with unique start positions ≥5 (**Table S2.2**).

DNA variants specific to the weeping pool and those common to both pools were considered genetically informative for mapping allele *weeping* (*W*), whereas variants specific to the standard pool were used as control (**Figure 2.1 vi**). Variants of allele frequency ranged from 15% to 80% were called heterozygous while those >80% were classified into homozygous (**Figure 2.1 vii**).

Mutant allele frequency (MAF) and density (MAFD) mapping

MAFD mapping is an adaptation of mutant allele frequency (MAF) mapping described previously (Korbinian Schneeberger, 2014) by integration of a second parameter-variant density.

It employs the weeping pool specific variants of segregation type <**l**m x mm> (Type-I), where allele '**l**' is a variant in relation to the reference genome and is assumed to be linked to allele *W*. In practice, the focus is on examining which genomic regions that might be situated with variants of average allele frequency close to 50% and how the variants were distributed along the genome (**Figure 2.1 viii**).

Allele frequency directional difference (AFDD) and density (AFDDD) mapping

AFDDD mapping explores two groups of variants common to both pools. The first group is of segregation type <**l**m x **ll**> (Type-II), where allele '**l**' is also a variant and assumed to be linked to weeping. The expected average variant allele frequency is 100% in the weeping pool and 50% in the standard pool in the *W* region. The second group of variants is of segregation type <**h**k x **h**k> (Type-III), where '**h**' is a variant and the first allele is linked to allele *W*. Under this scenario, the expected average allele frequency at the *W* locus is 75% in the weeping pool and 25% in the standard pool. The AFDD threshold in AFDDD mapping is AFDD≥30 percentage points. In this case, the goal is to look for the genomic regions that would situate with significantly more variants of average AFDD close to 50 percentage points (**Figure 2.1 viii**).

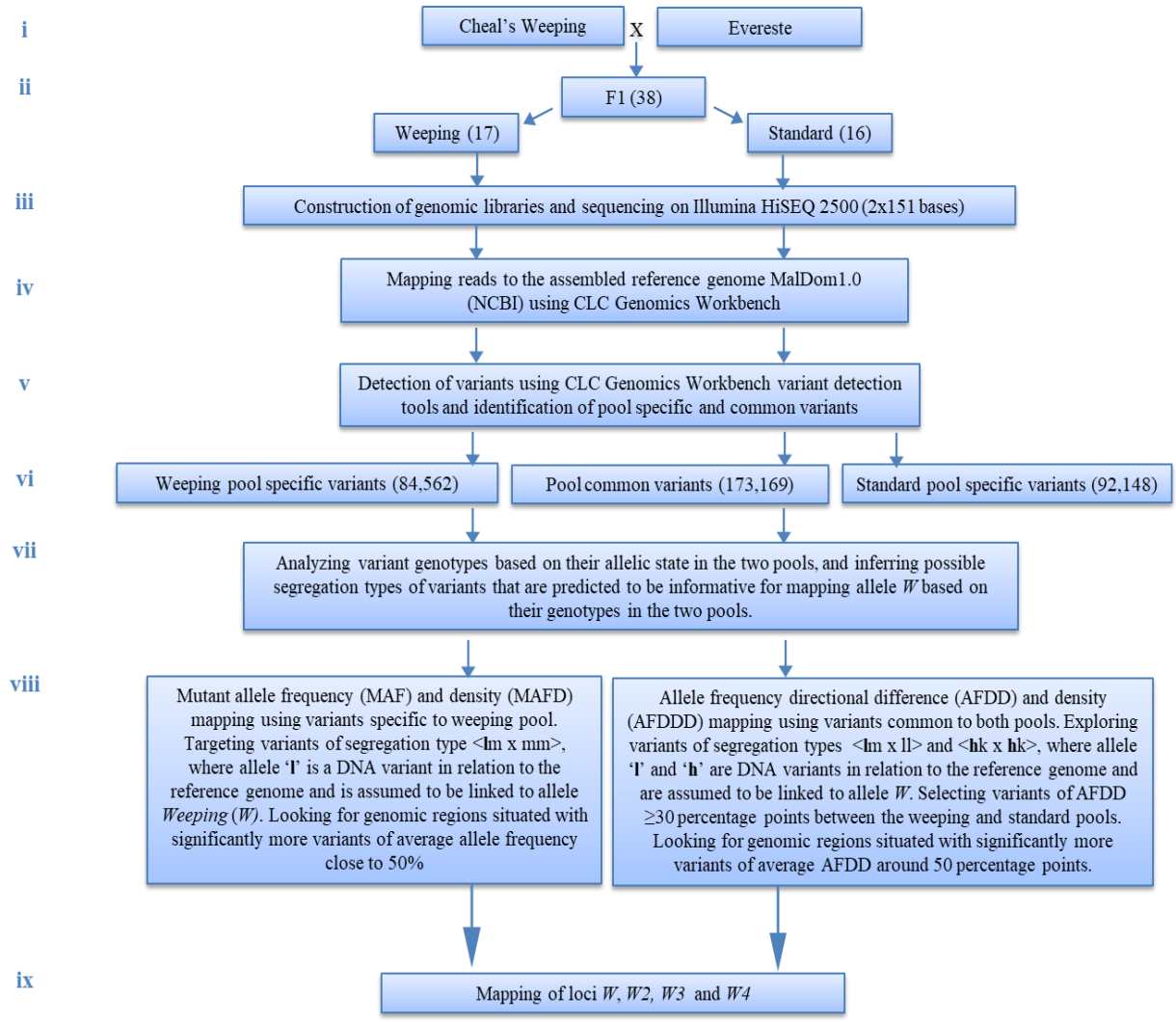


Figure 2.1. A flowchart illustrating the major steps in MAFD and AFDDD mappings of the weeping phenotype.

Standard score (z) test

Genome-wide distribution of DNA variants of the three segregation types <**l**m x mm>, <**l**m x **ll**> and <**h**k x **h**k> is assumed to be about even. In both MAFD and AFDDD mappings, if a genomic region is observed with a significant increase from the mean in variant density, the region is thought to be associated with the weeping phenotype. The significance test was conducted by standard score (z), which is calculated by the formula $z = (X - \mu) / \sigma$, where X =

observed variant density (variants/Mb); μ = mean variant density in the data set; σ = standard deviation of the mean (μ). The cut off is z = 2.6, p=0.01 (two-tailed confidence level).

Marker development

SSR markers were identified and developed from the apple reference genome sequence in the *W* and *W2* regions as previously described (Kenong Xu, Wang, & Brown, 2012). The primer sequence information and their approximate physical location in the genome were listed (**Table S2.3**). Polyacrylamide gel electrophoresis of SSR markers were conducted as detailed previously (Wang et al., 2012).

Sanger DNA sequencing

For confirmation of the variants of segregation types<**h**k x **h**k> and <**l**m x **ll**>, four genomic segments in the *W* region were PCR amplified with specifically designed PCR primers (**Table S2.3**) and the PCR products were sequenced directly by an ABI 3730XL DNA sequencer at the Cornell Genomics Facility Center.

RNA-seq and qRT-PCR analyses

Total RNA samples were isolated from actively growing shoot tip tissues of four weeping and four standard progeny individually from population 'Cheal's Weeping' × 'Evereste' using Qiagen Plant RNA Isolation Kit (Germantown, MD, USA). The Isolated RNA samples were pooled by phenotypes, forming a weeping and a standard RNA pool, respectively. Construction of RNA-seq libraries for the weeping pool and the standard pool were conducted similarly as described earlier (Bai et al., 2014). Single-end sequencing of read length 76 bp was performed on an Illumina NextSEQ 500 platform. RNA-seq reads were mapped to the improved or the latest

version of the apple reference transcriptomes (Bai et al., 2014) using CLC Genomics Workbench. Validation of RNA-seq analysis was performed by qRT-PCR assays on ten selected genes in the four weeping and four standard progeny. The qRT-PCR procedures were similar to what was described previously (El-Sharkawy, Liang, & Xu, 2015), and the primers, including those for the reference gene *MdActin*, were listed (**Table S2.4**).

BLAST-based dot matrix analysis

BLAST-based dot matrix analysis was performed using the BLAST tool for aligning two sequences, which is available at the NCBI website (https://www.ncbi.nlm.nih.gov/). The input of query sequences is limited to the regions of *W, W2, W3 and W4* in accessions CM001038.1 (Chr13), CM001035.1 (Chr10), CM001041.1 (Chr16), and CM001030.1 (Chr5) from the first version of the apple reference genome (Velasco et al., 2010), respectively. The input of subject sequences is correspondingly the entire sequences of chromosomes 13 (CM007879.1), 10 (CM007876.1), 16 (CM007882.1) and 5 (CM007871.1) from the new apple reference genome (Bai et al., 2014). The regions of similarity are visualized by the dot matrix tool available at the NCBI website as well.

Results

Segregation of weeping growth habit

In population 'Cheal's Weeping' × 'Evereste' of 38 individuals, 19 were scored standard (16) or standard-like (3) growth habit, 17 were weeping (10) and weeping-like (7), and two were intermediate (**Figure 2.2 A**). In population NY-051 × 'Louisa' of 140 seedling trees, 64 were scored standard (56) or standard-like (8) growth habit, 70 were weeping (56) and weeping-like

(14), and three were intermediate (**Figure 2.2 B**). The remaining three were dead or too weak for evaluation. In the NY-011 × NY-100 population, 22 individuals were standard (16) and standard-like (6), whereas 17 were observed as weeping (15) and weeping-like (2) (**Figure 2.2 C**). Chi-square tests (excluding the intermediates) showed that the segregation of weeping (-like) and standard (-like) growth habits fit the 1:1 ratio in all three populations (p=0.52-0.74), suggesting that the weeping phenotype is largely a dominant trait controlled by a major locus, presumably *W*. Thus parents 'Cheal's Weeping', 'Louisa' and NY-100 are of genotype *Ww* at the *W* locus, and 'Evereste', NY-051 and NY-011 are of genotype *ww*. The presence of individuals of less typical weeping and standard phenotype and intermediates in these populations suggests other modifying factors may exist.

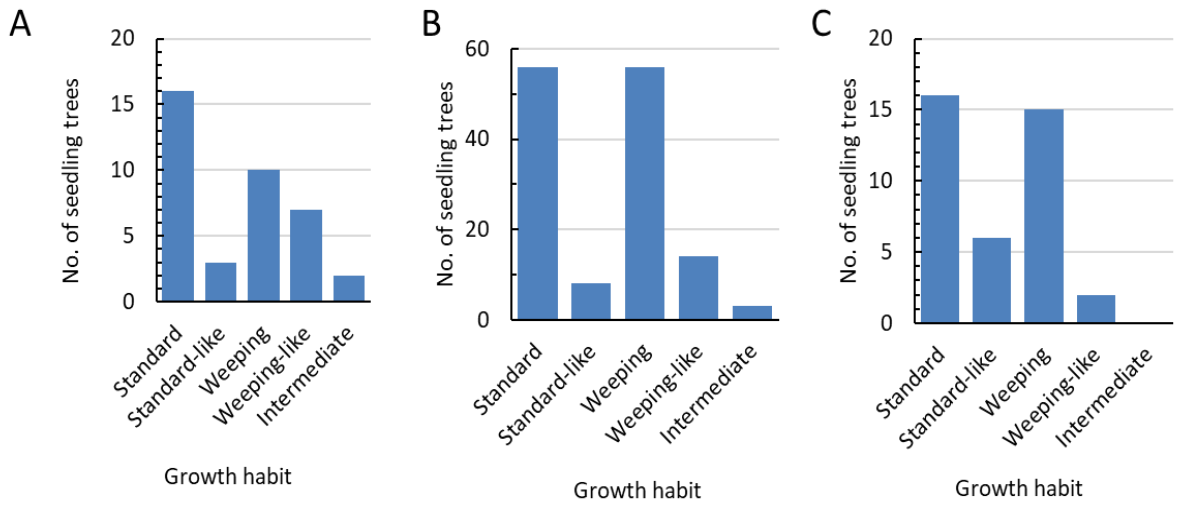


Figure 2.2. Phenotypic evaluation of growth habit in populations 'Cheal's Weeping' × 'Evereste' (**A**), NY-051 × 'Louisa' (**B**) and NY-011 × NY-100 (**C**) segregating for weeping phenotype. Chi-square tests (excluding the intermediates) showed that the segregation of weeping (-like) and standard (-like) growth habits fit the 1:1 ratio in all the three populations: 'Cheal's Weeping' × 'Evereste' (χ^2=0.1111, p=0.74), NY-051 × 'Louisa' (χ^2=0.2687, p=0.60), and NY-011 × NY-100 (χ^2= 0.4100, p=0.52).

Pooled genome sequencing analysis and identification of DNA variants

Illumina sequencing generated 140,742,316 and 157,357,078 paired-end raw reads (2x151bp) for the weeping and standard genome pools, respectively (NCBI accession SRR5099729). After removing 7,425,504 (5.0%) low quality reads in the weeping pool and 7,917,860 (5.3%) in the standard pool, the cleaned 133,316,812 (27.1 x the reference genome of 742.3Mb in weeping pool) and 149,439,218 (30.4 x in standard) reads (**Table S2.1**) were used for alignment against the reference genome. The mapped reads were 57,639,266 for weeping and 66,798,158 for standard, accounting for 43.2% and 44.7%, and covering 16.5x and 19.2x of the assembled reference genome (526.2Mb), respectively (**Table S2.1**).

Using the variant detection tool of CLC Genomics Workbench, a total of 2,700,059 variants in weeping and 2,946,289 in standard pools were detected. The number of variants of non-reference allele was 1,306,887 (SNV: 88.5%) and 1,380,503 (SNV: 87.8%) in weeping and standard pools, respectively (**Table S2.2**). Comparing the non-reference variants between the two pools identified 498,386 unique to the weeping pool and 573,589 unique to the standard, and 799,089 in common. To use more reliable variants, another set of filters were applied (**Table S2.2**), leading to 84,562 variants specific to the weeping pool and 92,148 specific to the standard, and 173,169 common to both pools, which constitute the primary datasets of variants for mapping the weeping trait (**Table S2.2**, **Figure 2.1 vi**). For an overview of these variants, the distributions according to their genotypes, allele frequencies and home chromosomes were shown (**Figure S2.2**, **Figure S2.3**).

Inferring segregation types of variants

'Cheal's Weeping' × 'Evereste' is a cross between two heterogeneously heterozygous diploid parents. In such crosses, at least six segregation types are possible for a given DNA variant if phase is not considered, including <ab x cd>, <ef x eg>, <hk x hk>, <lm x ll>, and <nn x np> and <qq x qq>, where each letter in the denotation stands for one of the four DNA bases (A, C, G and T) in SNVs, or an allele of other DNA variant types. To be informative for mapping allele *W* in pooled genome sequencing analysis, variants must have: 1) a heterozygous genotype in 'Cheal's Weeping'; 2) a dense coverage throughout the genome; and 3) a segregation type segregating for a unique allele frequency in the weeping pool or a large difference in allele frequency between the weeping and standard pools so that discrimination is possible. It is expected that segregation types <nn x np> and <qq x qq> will not be informative based on criterion 1). Since the most abundant variants are single nucleotide variants (SNVs) involving two alleles, variants of segregation types <ab x cd> and <ef x eg>, which segregate for four and three alleles, respectively, likely would be much less frequent. Therefore, the remaining two segregation types <hk x hk> and <lm x ll> are predicted to be informative.

To develop an effective approach for genetic mapping of the weeping phenotype, the genotypes of the 173,169 variants common to both pools were compared based on their allelic state observed in the weeping and standard pools (**Figure 2.1 vii**), leading to five genotype groups: G1: heterozygous in weeping / heterozygous in standard (He-W/He-S); G2: Ho-W/He-S; G3: He-W/Ho-S; G4: Ho-W/Ho-S; and G5: 'Complex' for those of complex genotypes involving four or three different DNA bases (i.e. three or two DNA variants in relation to the reference), presumably caused by segregation types <ab x cd> or <ef x eg> (**Table S2.5**, **Figure 2.3**, **Figure S2.4**). G1 is the largest group of 144,558 (83.5%) variants, whereas G2 and G3 groups of 5,353

(3.1%) and 2,104 (1.2%) variants, respectively. The G4 and G5 groups had 16,963 (9.8%) and 4,191 (2.4%) variants, respectively (**Table S2.5**). The variants specific to the standard pool fall into group G2 and those specific to the weeping into G3.

Inferring segregation types conceivably responsible for the observed G1-G5 identified at least 12 possible segregation types when the phase of variants was considered (**Table S2.5**, **Figure 2.3**, **Figure S2.4**). Further analysis concluded that only segregation types <**l**m x mm> (Type-I) for G3, <**l**m x **ll**> (Type-II) for G2, and <**h**k x **h**k> (Type-III) for G1 are informative for mapping of *W*, where the alleles at the first position are designated to be linked to weeping phenotype in the seed parent 'Cheal's Weeping' and those in bold are polymorphic variants in relation to the apple reference genome (**Table S2.5, Figure 2.3**, **Figure S2.4**). Obviously, Type-I variants are specific to the weeping pool, whereas Type-IIs and IIIs are common to both pools. An important common character of the three informative segregation types is that the variant allele frequencies in the weeping pool are higher than those in the standard pool by 50 percentage points, providing a practicably measurable directional (positive) difference in variant allele frequency between the weeping and standard pools (**Table S2.5**, **Figure 2.3**).

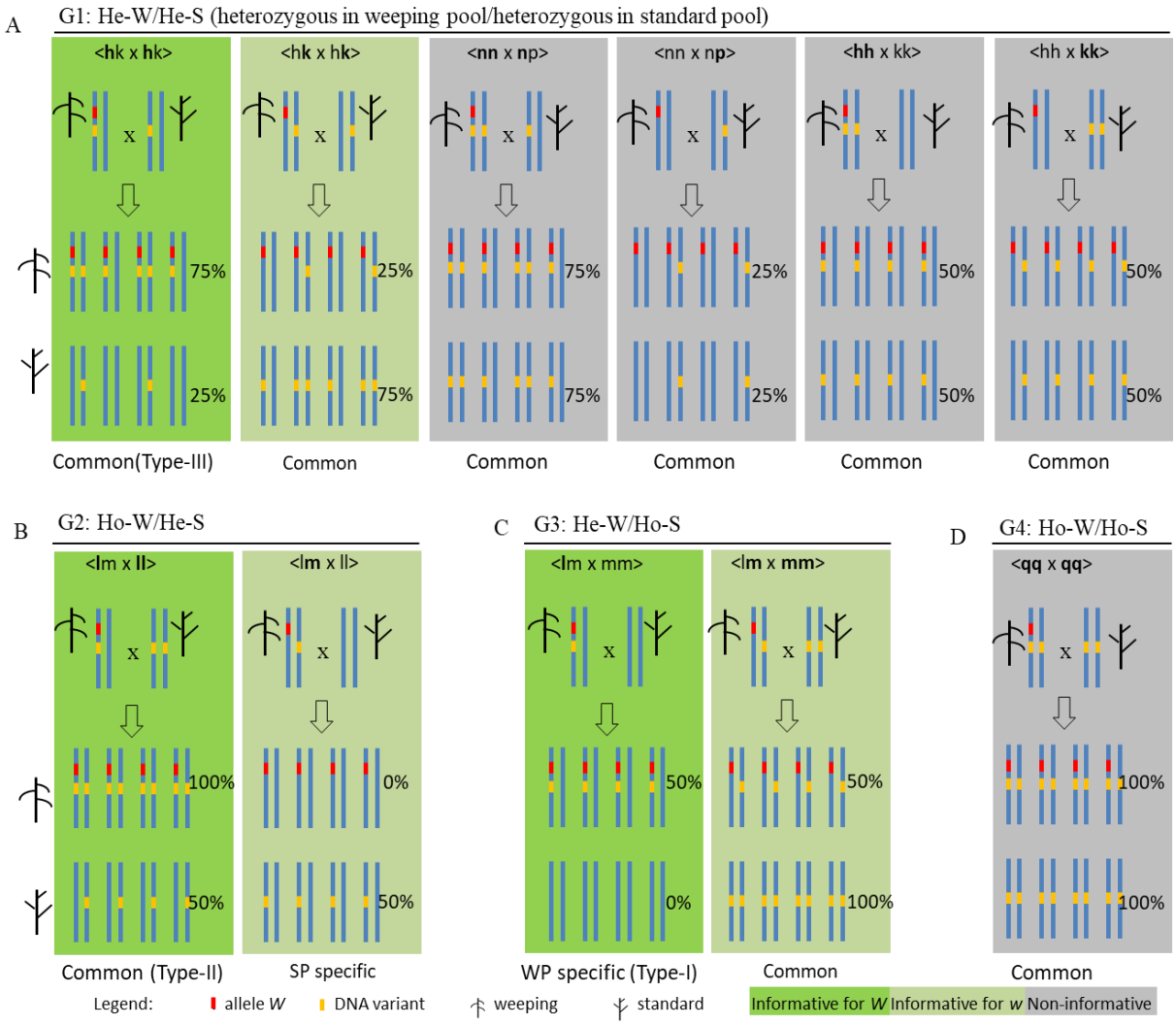


Figure 2.3. Schematics of the segregation of DNA variants linked to allele *W* in either phase under varying segregation types inferred for four of the five variant genotype groups-G1 (**A**), G2 (**B**), G3 (**C**) and G4 (**D**). Each segregation type is illustrated in a color filled rectangular, which includes the two parents at the top, four representative weeping progeny in the mid, and four standard progeny at the bottom. The long vertical lines in blue stand for the chromosomal segment harboring *W*. The red and orange short vertical lines represent allele *W* and DNA variants in relation to the reference genome, respectively. The tree-like drawings with up- and down-ward 'branches' indicate standard and weeping tree phenotypes, respectively. The expected allele frequency (%) of DNA variants in the weeping and standard pools is given accordingly. In each segregation type denotation, the allele at the first position is designated to be linked to weeping phenotype in the seed parent 'Cheal's Weeping' (e.g. letter 'l' in <**l**m x mm>), and those in bold are DNA variants in relation to the apple reference genome (e.g. letters 'l' in <**l**m x **ll**>). Segregation types informative for mapping allele *W* is shown in green rectangular (See Table S5 for more details). 'Common': variants common to both pools. SP: standard pool. WP: weeping pool.

Mapping of the weeping phenotype using weeping pool specific variants

Mapping of allele *W* was first conducted using the 84,562 weeping pool specific variants under the assumption that variants linked to the causal mutation for weeping phenotype are heterozygous in 'Cheal's Weeping' while homozygous in 'Evereste', i.e. following segregation type <**l**m x mm> (Type-I) Under this assumption, the average allele frequencies of these variants are anticipated to approach 50% in the *W* region (**Table S2.5**, **Figure 2.3 C**). To map allele *W*, the allele frequency data of the 84,562 variants were plotted against the apple reference genome (**Figure 2.4 A**). To facilitate visual inspection, their average frequencies were calculated in a moving window of 20 variants. The results showed that the moving average frequency mostly ranged from 30% to 40% throughout the genome, consistent with their frequency distributions (**Figure S2.3 A**). Interestingly, there is a most visible region of a moving average allele frequency around 50% as expected for Type-I variants, located on chromosome 13, suggesting that chromosome 13 putatively harbors the major locus *W*. In addition to chromosome 13, there appeared to be a number of regions on other chromosomes, such as 5 and 10, of allele frequencies around 50%, implicating uncertainties in allele frequency mapping.

To examine the uncertainties, the 18,604 variants of allele frequency close to 50% (40%-60%) were selected from the 84,562 weeping pool specific variants and were used to estimate variant density - the number of variants per million base pairs (Mb) DNA - throughout the genome (**Figure 2.4 B**). There were seven significant peaks of variant density, including the major peak on chromosome 13 (z=5.1, p=3.4E-06) and others on chromosomes 5, 8, 10, 12, 14 and 16. Given the number of putative regions identified, we sought other approaches to confirm the findings.

Since weeping represents a natural occurring mutation, this mapping strategy of using allele frequency of mutant pool specific variants together with their variant density is dubbed mutant allele frequency (MAF) and density (MAFD) mapping (**Figure 2.1 viii**).

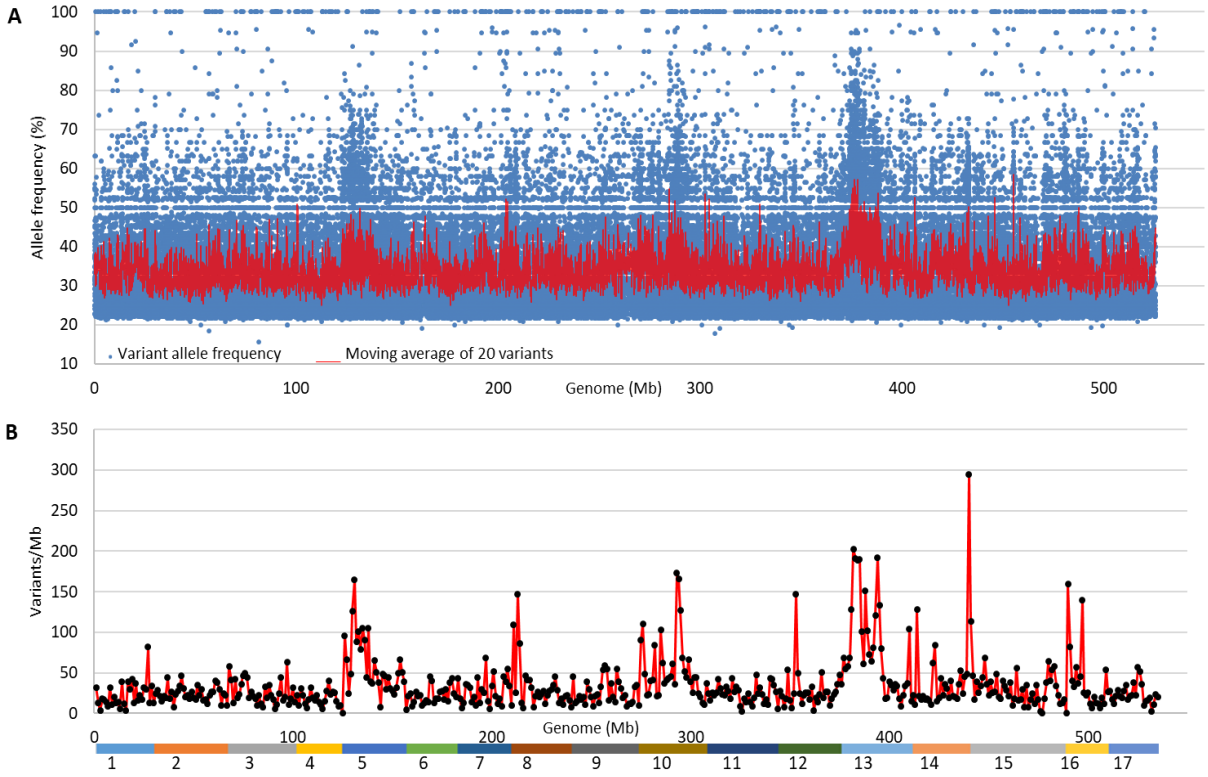


Figure 2.4. Distribution of allele frequency and density of variants specific to the weeping pool**.** (**A**) Distribution of allele frequency of 84,562 variants. (**B**) Distribution of density of 18,604 variants of allele frequency ranging from 40% to 60%. The color bar at the bottom represents the assembled reference genome of 17 chromosomes as numbered. Based on z-score test, significant variant density peaks were detected on seven chromosomes, including 5 (z=4.0, p=6.4E-05), 8 (z=3.4, p=6.7E-04), 10 (z=4.2, p=2.6E-05), 12 (z=3.4, p=6.7E-04), 13 (z=5.1, p=3.4E-06), 14 (z=7.9, p=0), and 16 (z=3.8, p=1.4E-04).

Mapping of the weeping phenotype using variants common to both pools

As shown earlier, variants common to both pools could be exploited for mapping based on Type-II and III variants. A positive 50-perecentage-point difference in allele frequency between the weeping and standard pools is expected for these two segregation types of variants.

With this understanding, variants common to both pools of an allele frequency directional (positive) difference (AFDD) ≥30 percentage points would be sufficiently inclusive and informative to map allele *W*. In total, 6,377 of the 173,169 variants were found with an AFDD≥30 percentage points. Plotting the 6,377 variants against the genome revealed that there are seven regions on chromosomes 1, 5, 6, 10, 13, 15 and 16 that show a moving AFDD around 50 percentage points (**Figure 2.5 A**).

The density distribution of the 6,377 variant of AFDD≥30 percentage points in the genome indicated that there were four significant variant density peaks located on chromosomes 13 (p=0), 10 (p=1.2E-08), 16 (p=1.1E-05), and 5 (p=3.7E-03) (**Figure 2.5 B**). Since they also were identified in MAFD (**Figure 2.4 B**) and AFDD (**Figure 2.5 A**) mappings, the four peak regions were concluded significantly associated with weeping.

A close look at the variant distribution for both AFDD and variant density revealed that the peaks covered:

1) a 7(5^{th}-12^{th})-Mb region on chromosome 13 (**Figure 2.6 A, B**), presumably including the weeping allele *W*,
2) a 4(18^{th}-22^{nd})-Mb region on chromosomes 10 (**Figure 2.6 C, D**), designated *W2*,
3) a first 2-Mb region on chromosome 16 (**Figure 2.6 E, F**), designated *W3*, and
4) a 3(4^{th}-7^{th})-Mb region on chromosome 5 (**Figure 2. 6 G, H**), designated *W4*

For convenience, such mapping processes that relies on DNA variants not only common to both pools, but also with AFDD ≥30 percentage points and density polarity towards the pooling selection targeted genomic regions is called allele frequency directional difference (AFDD) and density (AFDDD) mapping (**Figure 2.1 viii**).

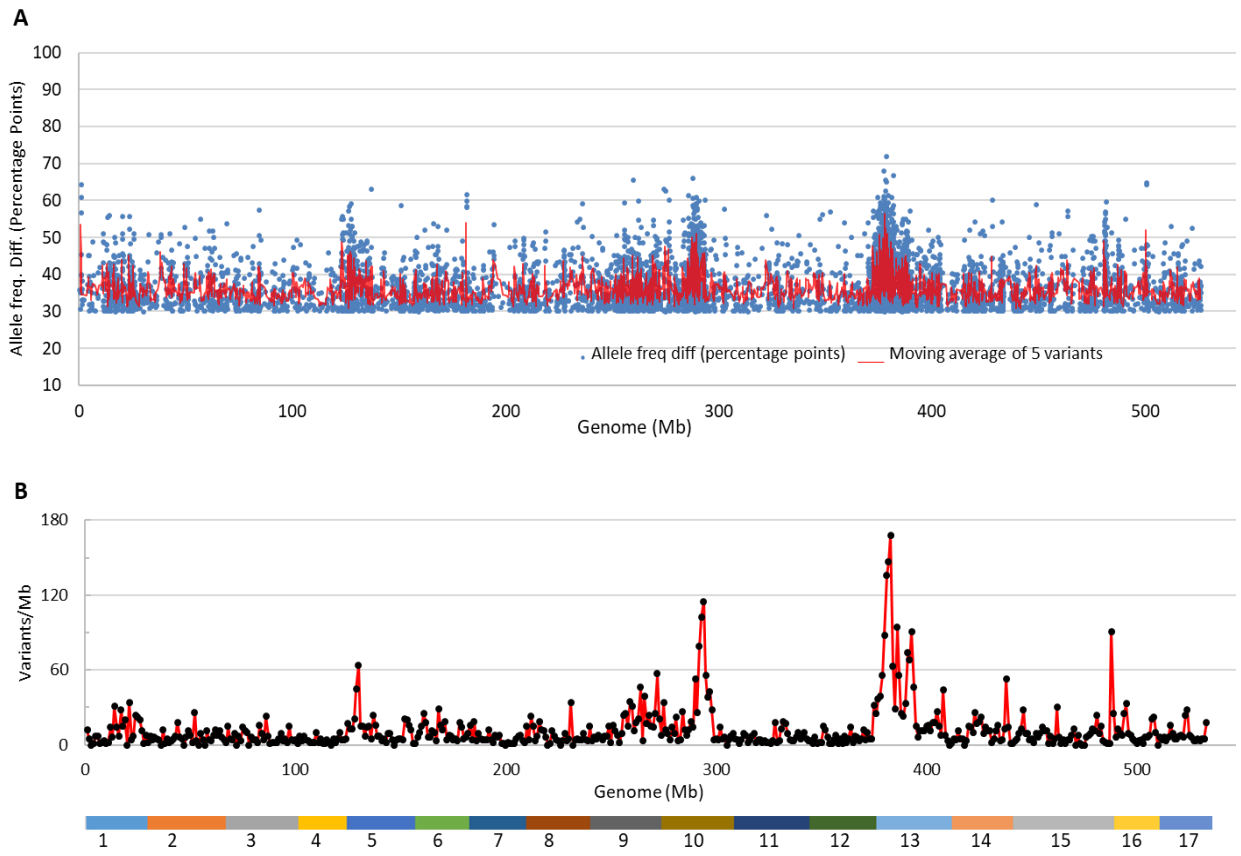


Figure 2.5. Distribution of allele frequency directional difference (AFDD) and density of variants common to both pools on the apple reference genome. (**A**) Distribution of AFDD of the 6,377 variants of AFDD≥30 percentage points between the weeping and standard pools. (**B**) Distribution of density of the 6,377 variants. The color bar at the bottom represents the assembled reference genome of 17 chromosomes as numbered. Significant variant density peaks were identified on chromosomes 13 (z=8.7, p=0), 10 (z=5.7, p=1.2E-08), 16 (z=4.4, p=1.1E-05), and 5 (z=2.9, p=3.7E-03).

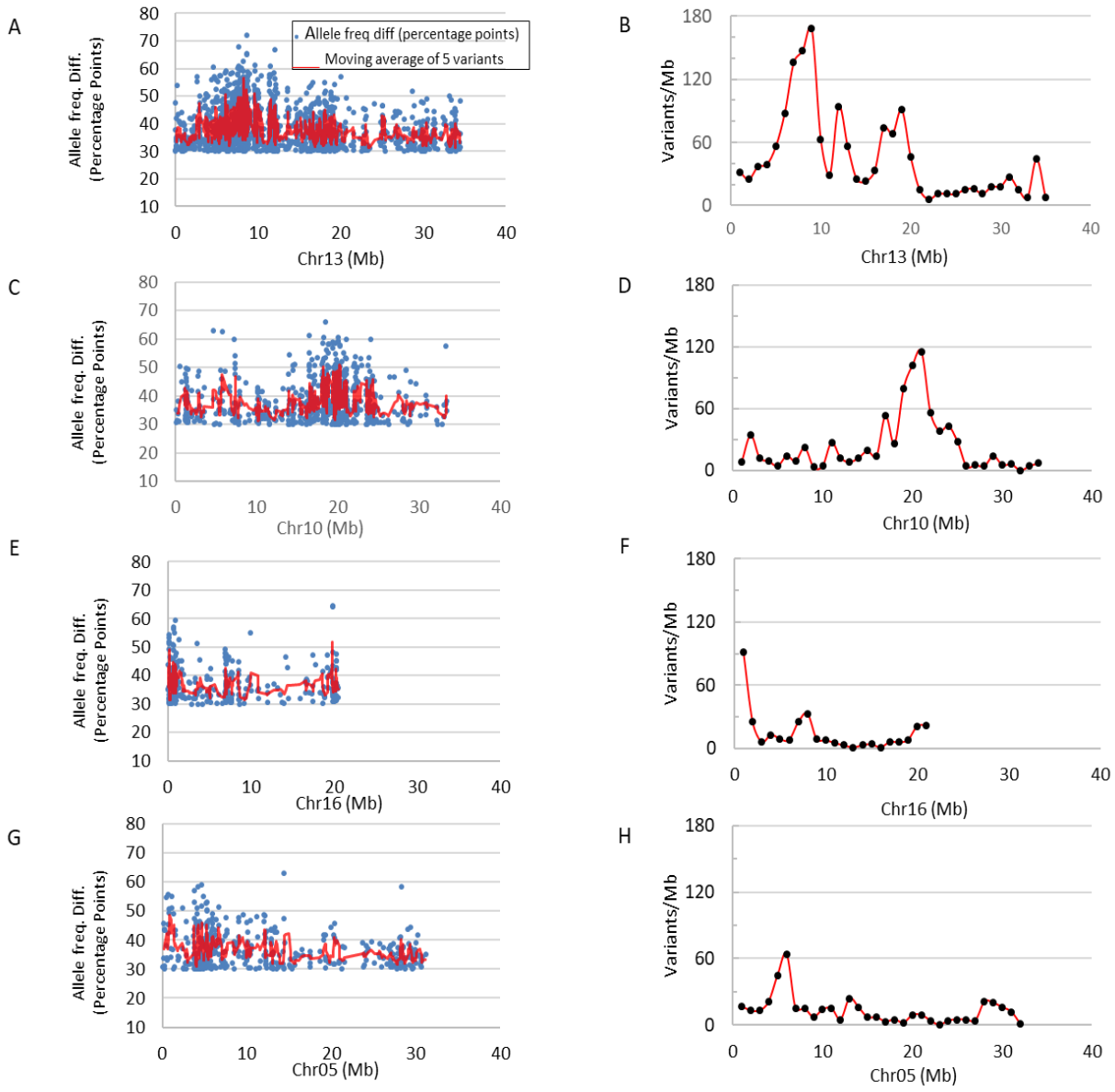


Figure 2.6. Distribution of allele frequency directional difference (AFDD) and density of variants with AFDD≥30 percentage points on chromosomes 13 (**A**, **B**), 10 (**C**, **D**), 16 (**E**, **F**) and 5 (**G**, **H**).

Evaluation of variant genotype groups in the two pools targeted by AFDDD mapping

To see what and how the variant genotype groups in the two pools (**Table S2.5, Figure 2.3**) were targeted in AFDDD mapping, their frequencies in the 6,377 variants of AFDD≥30 percentage points were analyzed at the levels of genome, chromosome 13 and the 7-Mb region of

W (**Figure 2.7 A-C**). Compared with the frequency of the 173,169 variants common to both pools in the five genotype groups (**Table S2.5**, **Figure 2.7 D**), AFDDD mapping clearly selected for variants in G2, against G1, G3 and G4, and neutrally for G5. It drastically increased the frequency in G2 from 3.1% (**Figure 2.7 D**) to 24.3% at the genome level (**Figure 2.7 A**), to 28.7% on chromosome 13 (**Figure 2.7 B**), and to 44.4% in the *W* region (**Figure 2.7 C).** Meanwhile, AFDDD mapping decreased the frequency in G2 from 83.5% (**Figure 2.7 D**) to 73.4% (**Figure 2.7 A**), 69.8% (**Figure 2.7 B**) and 53.4% (**Figure 2.7 C**) at the three levels, respectively. Since groups G2 and G1 accounted for 97.8% in the *W* region, and variants in the G2 is mostly, if not all, of segregation type <**l**m x **ll**> (Type-II), and only a part of those in G1 is of genotype segregation type <**h**k x **h**k> (Type-III), AFDDD mapping primarily targets Type-II and Type-III variants (**Table S2.5**, **Figure 2.3**, **Figure 2.7**).

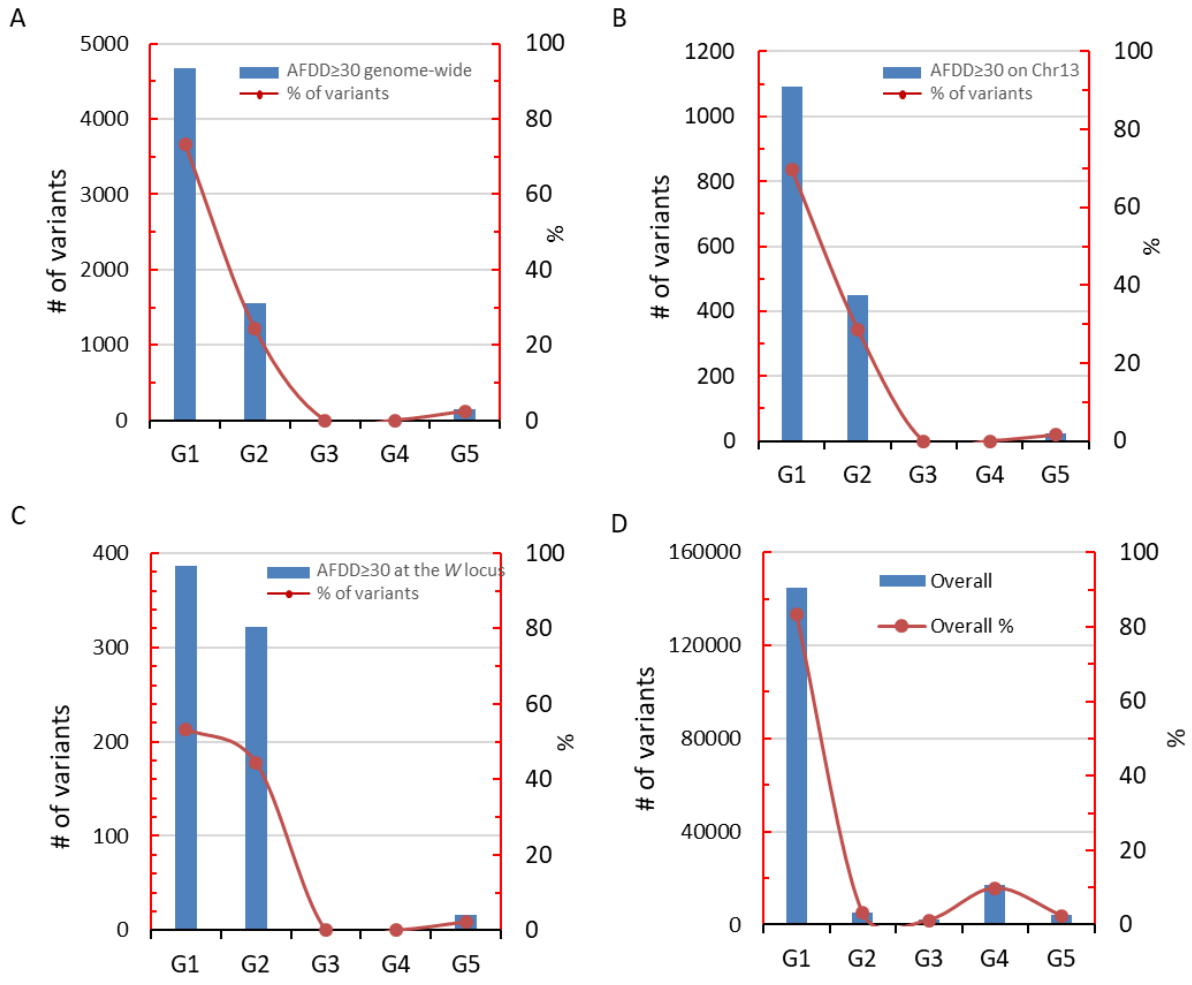


Figure 2.7. Assessing AFDDD mapping targeted variant genotype groups using the 6,377 variants of AFDD≥30 percentage points. (**A**-**C**) Number and frequency (%) of such variants observed in the five genotype groups at the genome scale (**A**), on chromosome 13 (**B**) and in the *W* region (**C**). (**D**) Number and frequency (%) of all the 173,169 variants common to both pools in the five genotype groups. G1: heterozygous in weeping / heterozygous in standard (He-W/He-S); G2: Ho-W/He-S; G3: He-W/Ho-S; G4: Ho-W/Ho-S; and G5: 'Complex'

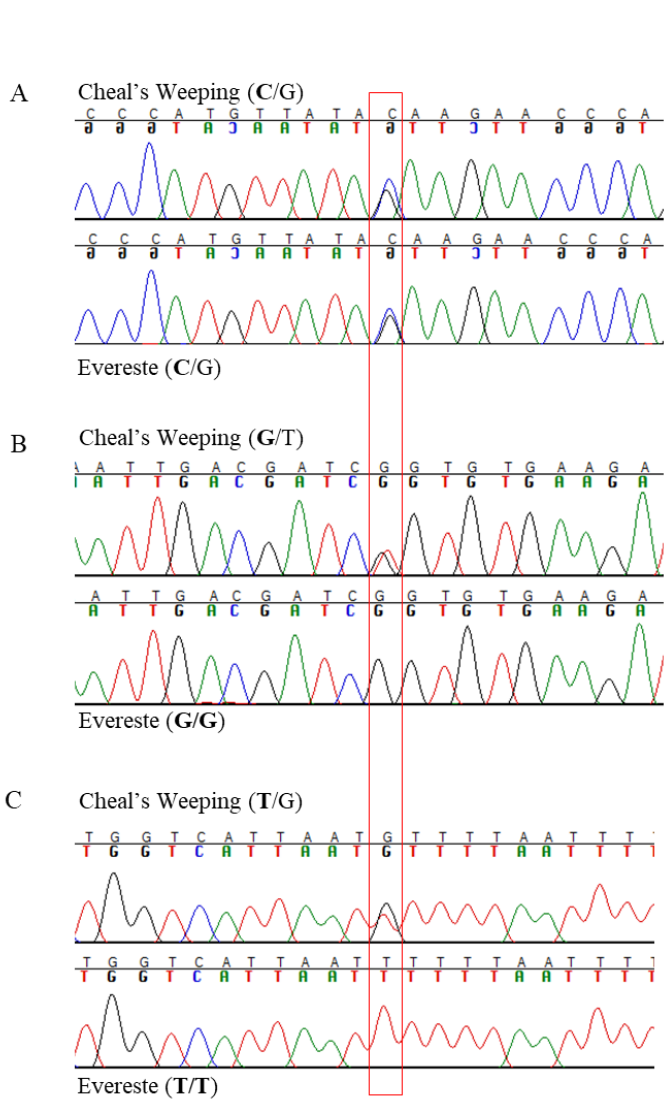
Analysis of AFDDD mapping contributing segregation types

To examine the contributing roles of Type-II and Type-III variants to AFDDD mapping, two sets of variants were selected in the weeping pool from the 173,169 variants common to both pools: one was the 15,425 variants of allele frequency ≥95%; the other was the 12,219 variants of allele frequency ranging from 70% to 80%. The assumptions are that under such selections in

the weeping pool, the expected responses of Type-II and Type-III variants would be a marked increase in their numbers in the *w* region in the standard pool, which should be characterized with allele frequency close to 50% and 25%, respectively. A detailed analysis using these two sets of variants confirmed the assumptions (**Figure S2.5**, **Figure S2.6**), highlighting their essential contributing roles in AFDDD mapping.

DNA evidence in support of AFDDD mapping

To directly confirm the presence of variants of segregation types <**l**m x **ll**> (Type-II) and <**h**k x **h**k> (Type-III), four genomic segments covering 14 putative variants (**Table S2.3**), including 12 for Type-II and two for Type-III in the *W* region were PCR amplified from the two parents (‘Cheal’s Weeping’ and ’Evereste’). Sanger DNA sequencing analysis of the PCR products demonstrated that all the 14 variants were confirmed to have the expected genotypes in the two parents (**Figure 2.8**, **Table S2.3**), providing physical evidence that variants of segregation types <**l**m x **ll**> and <**h**k x **h**k> are among those identified at the *W* locus by AFDDD mapping. Taken together, these data strongly support the role of Type-II and Type-III variants in AFDDD mapping.



	Genotype in reference genome	Genotype in W-pool and variant observed (and expected) frequency (%)	Genotype in S-pool and variant observed (and expected) frequency (%)
A	G/G	**C**/**C** **C**/G 69.6 (75.0)	G/**C** G/G 30.8 (25.0)
B	T/T	**G**/**G** **G**/**G** 100 (100)	T/**G** T/**G** 42.9 (50.0)
C	G/G	**T**/**T** **T**/**T** 100 (100)	G/**T** G/**T** 37.5 (50.0)

Figure 2.8. Chromatogram of DNA sequences of parents 'Cheal's Weeping' and 'Evereste' covering three SNVs (indicated by the red box) of segregation types <**h**k x **h**k> (**A**) and <lm x **ll**> (**B**, **C**) in the W region on chromosome 13. The SNV genotypes in the two parents, the reference genome and the weeping and standard pools are listed accordingly. (**A**) SNV at position 7,923,460 in gene LOC103452418. (**B**) SNV at position 8,209,678 in gene LOC103452141. (**C**) SNV at position 8,210,175 in gene LOC103452141. Letters in bold stand for DNA polymorphism (variant) in relation to the reference.

Confirmation of the mapping of locus W

To confirm the mapping of *W*, four SSR markers from the 7-Mb region of *W* (**Table S2.3**) were developed and analyzed. In population NY-051 × 'Louisa', based on the genotypic data and the progeny growth habit evaluations (**Figure 2.2 B**, **Figure 2.9 A**), markers SSR7641

and SSR8181 flank the *W* locus by three and one recombinants, respectively, from one side, and marker SSR9530 flanks *W* from the other side by two recombinants, delimiting the *W* locus within a 1.4 (8.1^{th}-9.5^{th})-Mb genomic region. Marker Ch13-8547 co-segregates with the weeping phenotype (intermediates discounted), confirming the mapping of *W*.

In population 'Cheal's Weeping' × 'Evereste', marker Ch13-8547 segregated 17 (*w*-linked allele):21 (*W*-linked-allele) for the two alleles from 'Cheal's Weeping' (Figure. 9B). Of the 17 progeny of the *w*-linked-allele, 14 were scored standard and three were standard-like, demonstrating a complete linkage to *w*. However, the 21 individuals of the *W*-linked allele were observed with a range of scores, including weeping (10), weeping-like (7), intermediate (2) and standard (2) (**Figure 2.9 B**).

In population NY-011 x NY-100, similar results were observed. Marker Ch13-8547 segregated with 16:23 for the *w*- and *W*-linked-alleles, respectively (**Figure 2.9 D**). The 16 progeny carrying the *w*-linked-allele showed normal growth habit, including 14 standard and two standard-like. Among the 23 individuals of the *W*-linked allele, 15 were noted as weeping, two as weeping-like, four as standard-like and two as standard (**Figure 2.9 D**). Therefore, these data confirmed the mapping of the major locus *W* on chromosome 13 in the three populations although the locus *W* could not explain the observations that eight progeny carrying the *W*- allele showed standard or standard-like phenotypes in populations 'Cheal's Weeping' × 'Evereste' and NY-011 x NY-100 (**Figure 2.9 B, D**).

Confirmation of the mapping of locus *W2*

For confirmation of the mapping of *W2*, three SSR markers (**Table S2.3**) from the *W2* region were developed and evaluated. In population NY-051 × 'Louisa',-markers Ch10-19768

and Ch10-20017 segregated normally, but did not show significant association with the weeping phenotype (data not shown), suggesting that mapping of *W2* could not be confirmed.

In population 'Cheal's Weeping' × 'Evereste', the segregation of markers Ch10-19768 and Ch10-20017 showed a significant association with the weeping phenotypes. For example, marker Ch10-20017 segregated the *w*-linked allele in 21 individuals and the *W*-linked allele in 17, fitting the 1:1 ratio (χ^2= 0.237, p=0.626) as expected (**Figure 2.9 C**). Among the 21 progeny of the *w*-linked allele, the ratio between weeping (-like) and standard (-like) was 6:15, significantly skewed towards the standard phenotype from the expected 1:1 ratio (χ^2=3.86, p=0.049). In contrast, a considerable skewness towards the weeping phenotype in the 17 progeny carrying the *W*-linked allele was observed (χ^2=3.27, p=0.071) as there were 11 weeping (-like) and only four standard (-like) individuals (**Figure 2.9 C**). The remaining two were intermediate and were not counted.

In population NY-011 x NY-100, markers Ch10-20017 and Ch10-20761 (**Table S2.3**) were informative for the weeping associated allele from 'Red Jade'. Unlike what was observed in population 'Cheal's Weeping' × 'Evereste', the segregation of marker Ch10-20017 was significantly distorted from 1:1 (χ^2=10.26, p=1.36 $\times 10^{-3}$) as its *W*- and *w*- linked alleles were observed in 30 and 9 progeny, respectively (**Figure 2.9 E**). A close look revealed that the marker segregated 15:7 for the *W*- and *w*- linked alleles in the 22 progeny of standard (-like) phenotype, fitting the ratio 1:1 (χ^2= 2.227, p=0.136). However, the marker segregated 15:2 for in the 17 progeny of weeping (-like) phenotype, significantly skewed towards weeping (χ^2=8.471, p=3.60 $\times 10^{-3}$), suggesting a significant linkage between marker Ch10-20017 and the weeping phenotype, confirming again the mapping of *W2*.

Genetic interactions between *W* and *W2*

Investigating the genetic interactions between the *W* and *W2* alleles indicated that allele *W* was required for the weeping phenotype as there were no weeping individuals scored when it was absent in genotypes *w*-*w2* and *w*-*W2* (**Figure 2.9 F-G**) in the two populations. Comparing the number of individuals of typical weeping phenotype in genotype *W*-*W2* with that in genotype *W*-*w2*, i.e. 8/14 (57.1%) vs 2/7 (28.6%) in 'Cheal's Weeping' × 'Evereste', and 13/18 (72.2%) vs 2/5 (40.0%) in NY-011 x NY-100, suggested that the presence of allele *W2* increased the penetrance of weeping phenotype from allele *W* (**Figure 2.9 F-G**). Notably, one of the two *W*-carrying progeny of standard phenotype was genotyped as *W*-*w2* in 'Cheal's Weeping' × 'Evereste', providing a likely cause for the observation although the other remains to be explained due to its genotype *W*-*W2*. The *W*-*w2* genotype appeared to be also responsible for the observation that there were three *W*-carrying progeny of standard-like phenotype in population NY-011 x NY-100 (**Figure 2.9 G**).

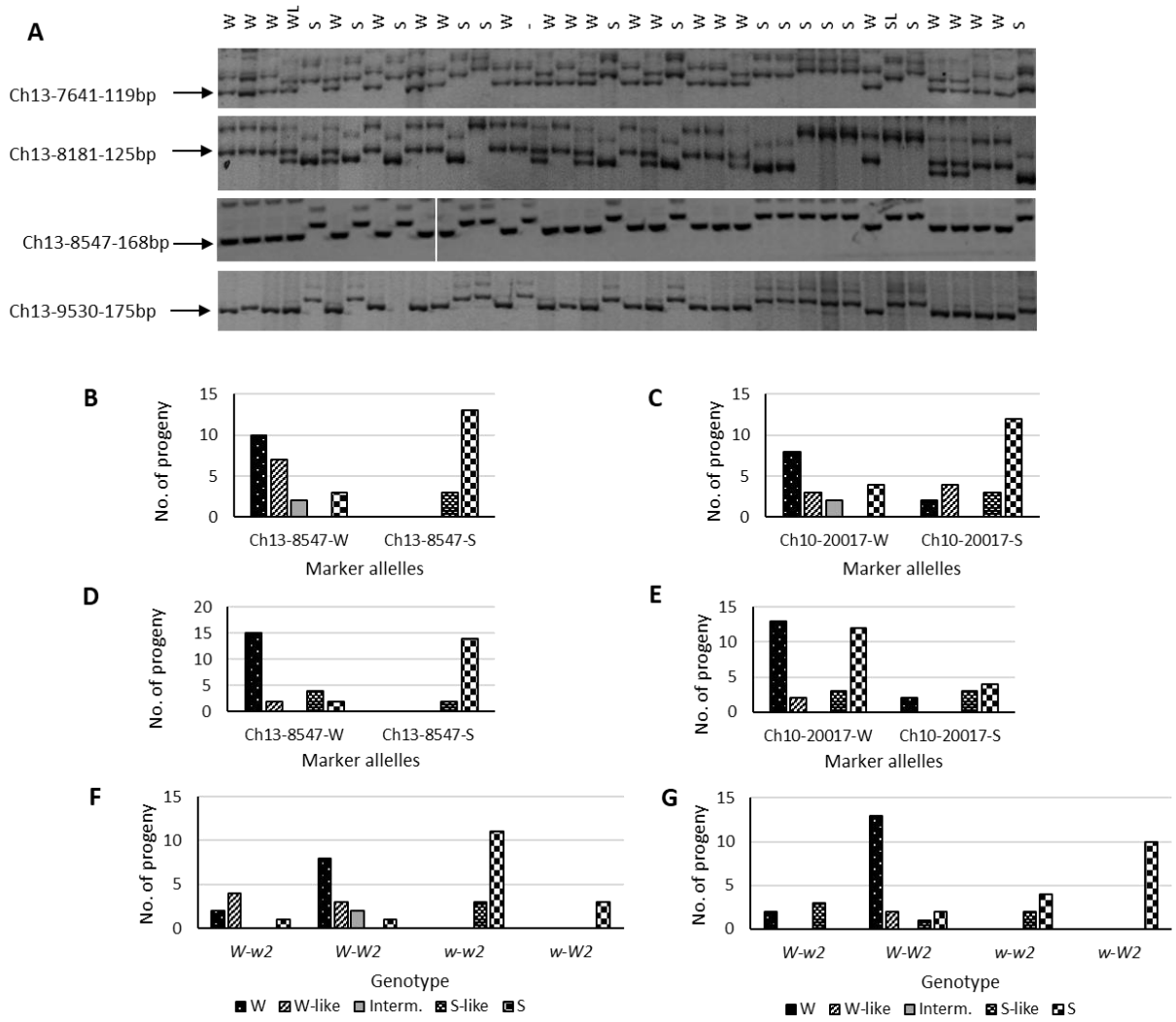


Figure 2.9. Confirmation of mapping of loci *W* and *W2*. (**A**) Analysis of four SSR markers Ch13-7641, Ch13-8181, Ch13-8547 and Ch13-9530 from the *W* region on chromosome 13 in population NY-051 × 'Louisa'. The pictures show the markers' polyacrylamide gel electrophoresis profile in 38 of the 140 individuals. The SSR bands Ch13-7641-119bp, Ch13-8181-125bp, Ch13-8547-168bp (the vertical line between lanes 10 and 11 in panel 3 indicates that this marker was run in two gels) and Ch13-9530-175bp of 'Louisa' origin and linked to the weeping phenotype are indicated with an arrow. W: weeping; WL: weeping-like; S: standard; SL: standard-like. -: seedling tree was dead before phenotyping. (**B-E**) Weeping trait association of SSR markers Ch13-8547 (in the *W* region) and Ch10-20017 (in the *W2* region) in populations 'Cheal's Weeping' × 'Evereste' (**B**, **C**) and NY-011 × NY-100 (**D**, **E**). Marker alleles linked to weeping and standard are suffixed with '-W' and '-S', respectively. (**F**, **G**) Effect of genetic interactions between the alleles of *W* and those of *W2* (deduced from marker alleles Ch13-8547-W and Ch10-20017-W, respectively) on expressivity of the weeping phenotype in populations 'Cheal's Weeping' × 'Evereste' (**F**) and NY-011 × NY-100 (**G**).

Identification of differentially expressed genes (DEGs) in the W and W2 regions

Based on the first version (V1) of the apple reference genome (Velasco et al., 2010), there are 153 and 368 genes or transcribed sequences in the *W* (1.4 Mb) and *W2* (4 Mb) regions. To examine their expression patterns, an RNA-seq analysis was performed using actively growing shoot tip tissues from four pooled weeping and four pooled standard progeny in 'Cheals Weeping' x 'Evereste'. A total of 43.2 million raw reads were obtained for the weeping and 59.2 million for the standard pools (**Table S2.6)** (NCBI accession SRR5760456). The RNA-seq analysis, which was validated by qRT-PCR testing on ten selected genes (Figure S7), identified 79 genes expressed (RPKM ≥1.0 in at least one of the RNA-seq pools) in the *W* region (Table S7) and 199 in the *W2* region (**Table S2.8**). There are three DEGs (p_{FDR}<0.05) in the *W* region, including MDP0000928608 (M928608) and M534197 (both glucuronoxylan 4-O-methyltransferase (GXMT)-like genes), and M160372 (a rubber elongation factor-like gene). In the *W2* region, five genes were expressed differentially, including G103289 and G104254 (both TMV resistance protein N-like genes), G102554 (an AUX/IAA7.1-like gene), M142356 (a chloroplastic decapping nuclease DXO-like gene), and M819881 (an E3 ubiquitin-protein ligase 3-like gene), where 'G######' stands for novel transcripts (Bai et al., 2014).

In the latest version (V2) of the apple reference genome (Bai et al., 2014), the *W* and *W2-W4* regions were all found on their corresponding chromosomes as those in V1 (Velasco et al., 2010) according to a BLAST-based dot matrix analysis although their chromosomal nucleotide coordinates were different (**Figure S2.8 A-D**). The *W* region between markers SSR8181 and SSR9530 was found in less than 1-Mb segment from 8.650th Mb to 9.635th Mb on chromosome 13 (**Figure S2.8 A**) that contains 72 predicted genes, whereas the *W2* region was determined to span over 3.5 Mb from 27.5th Mb to 31.0th Mb on chromosome 10 (**Figure S2.8 B**), where 216

genes are annotated. RNA-seq analysis (**Table S2.6**) showed that 60 of the 72 genes in the *W* region (**Table S2.9**) were expressed, of which eight were DEGs, including three (MD13G1118800, MD13G1119900 and MD13G1120100) that correspond to the three DEGs identified in V1 (**Table S2.7**, **Table S2.9**, **Table S2.11**). The remaining five DEGs includes three (MD13G1119100-photosystem I light harvesting complex gene 2; MD13G1121000-nuclear transport factor 2A; and MD13G11233000-BCL-2-associated athanogene 5) that were non-DEGs in V1 and two (MD13G11271000-jasmonate-zim-domain protein 10; and MD13G1127300-acyl-CoA oxidase 1) that were immediately outside the *W* region in V1 (**Table S2.11**). In the *W2* region, 151 of the 216 genes were expressed and six of them were DEGs (**Table S2.10**). Three (MD10G1192900, MD10G1199900 and MD10G1202900) of the six DEGs were identified in V1 as well, equivalent to G104254, M142356 and M819881, respectively (**Table S2.11**). The other three DEGs include MD10G1196900 (glutathione S-transferase TAU 19) and MD10G1203300 (ribosomal protein S5/Elongation factor G/III/V family protein) that were non-DEGs in V1, and MD10G1196600 (glutathione S-transferase TAU 25) that was not annotated in V1.

Discussion

Challenges in pooled genome sequencing-based genetic mapping in Malus and MAFD and AFDDD mappings

DNA variants of segregation type <**l**m x mm> (Type-I) in the weeping (mutant) pool are an obvious target to be exploited in genetic mapping studies involving an F_1 population derived from two parents of heterogeneously heterozygous genome such as ‘Cheal’s Weeping’ x ‘Evereste’. We anticipated that application of MAFD mapping using the weeping pool specific

variants would lead to a relatively 'clean' mapping of the weeping phenotype as it uses an approach analogous to a combination of MAF mapping (Korbinian Schneeberger, 2014) and variant density estimates similar to what was described previously (Minevich et al., 2012; Zuryn et al., 2010). This was true for mapping the major weeping locus *W*, but the allele frequency-based approach was found to have generated a number of other regions putatively associated with weeping phenotype (**Figure 2.4 A**). The variant density-based approach appeared to provide improved resolution, but it reported the *W* locus along with six other genomic regions significantly associated with weeping (**Figure 2.4 B**), raising questions about the utility of such a straightforward adaptation of existing mapping strategies to the most commonly used Type-I variants.

AFDDD mapping is an approach developed to address the uncertainty issue encountered in MAFD mapping by unlocking genetic information from DNA variants of other hidden segregation types for mapping. It differs from MAFD mapping in the follow ways: 1) AFDDD mapping focuses on variants of segregation types <**l**m x **ll**> (Type-II) and <**h**k x **h**k> (Type-III) rather than <**l**m x mm> (Type-I); 2) it uses the common variants between the pools rather than the weeping pool specific variants; 3) it examines variant allele frequency directional differences (AFDD) between the weeping and standard pools rather than their original allele frequency. Similar to MAFD mapping but distinct from other approaches, such as delta SNP index mapping (Fekih et al., 2013; Takagi et al., 2013), AFDDD mapping also emphasizes variant density by examining how variants were distributed preferentially in the pooling selection targeted regions than in the other non-linked regions. As such, AFDDD mapping was found to be an effective approach for identifying regions of major or minor effect, similar to quantitative trait loci (QTLs), on weeping in *Malus* (**Figure 2.5**, **Figure 2.6**).

An advantage in AFDDD mapping is that the variants common to both pools are of higher accuracy than the pool specific variants. This is likely due to the fact that common variants were each recognized twice, once in the weeping pool and the other in the standard, whereas the pool specific variants were called only once in one of the two pools. It is common to see genomic regions of uneven sequencing coverage between the pools. If a variant is called from a region in one pool, but not called in the same region in the other pool due to lower coverage, the variant would be falsely identified as specific to the first pool. In search for genetic variations between apple cultivars 'Gala' and 'Blondee', a yellow fruit somatic mutation of 'Gala' using RNA-seq data, 'Blondee' specific variants were proven to be false positive due to uneven coverage (El-Sharkawy et al., 2015). Since the pool common variants must be called in both pools, such false positive variants can be eliminated. In addition, it has been challenging in variant calling in NGS data analysis (Huang et al., 2015; Ribeiro et al., 2015).

A drawback in AFDDD mapping is that the causal mutations cannot be readily identified using the variants common to both pools for a dominant trait. This drawback of AFDDD mapping, however, underscores a major advantage in MAFD mapping as the causal mutations would be present in the mutant pool specific variants. Therefore, for improving the mapping results and for identifying candidate causal mutations, simultaneously using AFDDD mapping together with MAFD mapping provides an improved strategy for pooled genome sequencing data analysis in *Malus*. An alternative approach to address the drawback in AFDDD mapping is to conduct another round of AFDDD mapping by copying a small fraction of randomly selected reads in mutant pool, e.g. equivalent to 10-15% of reads in wild type pool, into the wild type pool before variant calling. It is expected that such manipulation in reads pooling would make

the causal variants detectable in the modified wild type pool as long as a lower variant read count threshold (e.g. 10-15%) is used in variant calling.

Allelic distance (AD) is defined as the sum of absolute differences in the allele (variant) counts within a given region between two pools. AD mapping has been shown to be useful for mapping mutation using pooled genome sequencing data (Korbinian Schneeberger, 2014). The genetic presumptions are that the allelic distance between two pools of contrasting phenotype would be the largest in the causal region as the alleles of the mutant background are expected to be more concentrated there than in other non-linked regions (Korbinian Schneeberger, 2014). Since AD mapping relies on counting the number of variants, including those from pool-specific and pool-common variants, this likely would make it less effective in heterogeneously heterozygous species. In this study, if AD mapping were used, the input variants would include those both specific and common to the pools, i.e. 349,859 (sum of 84,562-weeping pool specific, 92,148-standard pool specific and 173,169-common to both pools, Figure. 1vi). To be effective, the 173,169 common variants would be excluded to avoid the results being obscured. This would virtually be equivalent to conducting the variant density assays between the two pools. Since the *w* region on chromosome 13 also showed a significantly higher variant density (data not shown), using AD mapping would have rendered the major locus *W* undetectable.

Genetic basis of AFDDD mapping

Through a detailed analysis of allele frequency and genotype of variants in the weeping and standard pools, a series of hypothetical variant segregation types were inferred, leading to identification of three that were used for mapping, including <**l**m x mm> suitable for the weeping pool specific variants, and <**h**k x **h**k> and <**l**m x **ll**> for variants common to both pools

(**Table S2.5** and **Figure 2.3 A-C).** Although DNA variants of segregation type <**l**m x mm> (Type-I) are commonly used, variants of segregation types <**l**m x **ll**> (Type-II) and <**h**k x **h**k> (Type-III) have not been employed in pooled genome sequencing studies in out-crossing woody species. Successful mapping of loci *W, W2*, *W3* and *W4* based on Type-II and Type-III variants, and the DNA sequence confirmation of 14 such variants in both weeping and non-weeping parents 'Cheal's Weeping' and 'Evereste' (**Figure 2.8**) suggested that they represent a unique source of DNA variants that can be exploited in NGS based pooled genome sequencing analysis. Indeed, among the 173,169 variants common to both pools, 5,353 (3.1%) are type-II variants, and Type-III would be even more (**Table S2.5**, **Figure 2.7**). In addition, this study has also demonstrated such variants are present throughout the genome and were readily identifiable and selectable through genome pooling (**Figure 2.7**, **Figure S2.5**, **Figure S2.6**).

Comparing allele frequencies is necessary in all studies involving analysis of pooled DNA samples (Shaw, Carrasquillo, Kashuk, Puffenberger, & Chakravarti, 1998). In classic BSA (Giovannoni et al., 1991; Michelmore et al., 1991), identification of markers linked to a trait of interest is accomplished by analyzing the marker allele frequency difference between two groups of individuals forming the two pools and a level of difference at 100% is typically pursued (K Xu & Mackill, 1996; K Xu, Xu, Ronald, & Mackill, 2000). In pooled genome sequencing based mapping studies, variant allele frequency differences have also been exploited, such as delta SNP index mapping (Fekih et al., 2013; Takagi et al., 2013). In AFDDD mapping, the optimal AFDD level is 50 percentage points between the weeping and standard pools as determined by Type-II and Type-III variants. In this study, the threshold of AFDD ≥30 percentage points was chosen to accommodate deviations. Such magnitude of AFDD is likely the underlying reason for a 'clean'

mapping of the weeping phenotype using variant density, including the elimination of the highest density peak on chromosome 14 identified using Type-I variants (**Figure 2.4 B**).

Variant allele density is likely a parameter more important in AFDDD mapping than in MAFD as it performed better for mapping the weeping phenotype. Due to the inherent selection during genome pooling and local physical linkage, the causal mutation region is inevitably to have a higher density of variants than in the unlinked region in the pools (Minevich et al., 2012; K. Schneeberger et al., 2009; Zuryn et al., 2010), explaining the effectiveness of variant density as a mapping parameter. Interestingly, a recent study (Jensen, Fazio, Altman, Praul, & McNellis, 2014) in an apple rootstock segregation population reported that most of the differentially expressed genes associated with resistance to powdery mildew and woolly apple aphid were clustered each case in a 9-10Mb region. However, the underlying genetic basis for such observations likely differs from that for the MAFD and AFDDD mappings.

Genetic control of weeping in Malus

The weeping phenotype in other woody species is controlled mostly by a single recessive allele, such as *pl* (Dirlewanger & Bodo, 1994) or *we* (Chaparro, Werner, O'Malley, & Sederoff, 1994) in peach (*Prunus persica*), *pl* (Zhang et al., 2015) in mei or Japanese apricot (*Prunus mume*), and *wp1* (Roberts, Werner, Wadl, & Trigiano, 2015) in eastern redbud (*Cercis canadensis*). In the case of mei, some modifier genes were likely involved in how the weeping phenotype is expressed in addition to *pl* (Zhang et al., 2015). Non-allelic weeping alleles were reported as well. For example, the weeping allele *wp1* is non-allelic to another weeping phenotype in eastern redbud (Roberts et al., 2015).

Using NGS-based MAFD and AFDDD mapping approaches, the present report uncovered four chromosomal regions *W* (chr13), *W2* (chr10), *W3* (chr16) and *W4* (chr5) (**Figure**

2.6) that are significantly associated with the weeping phenotype inherited from 'Cheal's Weeping'. This appeared to be consistent with what was suggested in another study using 'Red Jade' as a weeping parent (Just, 2001), where a number of progeny of intermediate phenotype were documented. Together with the independent confirmation of loci *W* and *W2* and the investigation into their genetic interactions (**Figure 2.9**), these findings provided important insight into the genetic control of weeping in *Malus*. The *W* locus is clearly of the most significant influence on weeping. For this reason, it is regarded the same locus for the major dominant gene *W* previously reported for *M baccata* 'Gracilis' (Alston et al., 2000; Susan K. Brown, 1992; Sampson & Cambron, 1965). Allele *W2* has been shown to play an important role in the weeping trait expressivity under the background of allele *W* in populations 'Cheal's Weeping' × 'Evereste' and NY-011 × NY-100, but a non-detectable role in population NY-051 × 'Louisa'. This indicates that 'Louisa' likely differs from 'Cheal's Weeping' and 'Red Jade' in the genetic mechanism responsible for the weeping phenotype expressivity, suggesting that 'Louisa' is distinct while 'Cheal's Weeping' and 'Red Jade' are related each other.

A study of the relatedness in a diverse set of *Malus* weeping accessions, based on genotypic data from seven SSR markers, showed that weeping accessions in crabapple were clustered into two clades: one is the *M. prunifolia* 'Pendula' and its descendants, and the other is the 'Hyvingiensis' group (Lindén & Iwarsson, 2014). The study also reported that apple cultivar 'Elise Rathke' accessions formed their own group distinct from all weeping accessions in crabapples. An important future study would be to understand if the *W* locus and any of the three other loci *W2, W3* and *W4* are also the key genetic factors determining their weeping phenotype.

Identification of the causal mutation is an important goal in pooled genome sequencing studies. Successful identification of causal mutations has been documented (Huo et al., 2016;

Petit et al., 2016; Schierenbeck et al., 2015). In this study, three or eight genes in the *W* region and five or six genes in the *W2* region were expressed differentially depending on the versions of the apple reference gnome used (Tables S7-S11). In the *W* region, the two GXMT like genes M534197 (MD13G1120100) and M160372 (MD13G1119900) are of more interest and had over two fold reduced expression in weeping compared to standard. It is known that GXMT-like genes function in biosynthesis of the hemicellulose 4-O-methyl glucuronoxylan, a major component comprising the secondary cell walls in dicots (Urbanowicz et al., 2012), and mutations in xylan synthesis genes often lead to unusual growth of plants (Carpita & McCann, 2015). Among the five or six DEGs in the *W2* region, the *AUX/IAA7.1*-like gene G102554 (MD10G1192900) appeared to fit well the role of *W2* and was upregulated in weeping, compared to standard. Several characterized *AUX/IAA* like genes in *Arabidopsis* were shown to act as repressor of auxin-inducible gene expression and to play roles in the control of gravitropic growth and development in light-grown seedlings (Sato, Sasaki, Matsuzaki, & Yamamoto, 2014, 2015; Yu et al., 2013).

The other genes of putatively regulatory roles in plant growth but not expressed differentially in statistics between weeping and standard progeny may not be ruled out. In the *W* region, these genes, for example, may include a LAZY1-like gene M374900 (MD13G1122400) and a transcription factor TCP20-like gene M254069 (MD13G1122900) as LAZY1s are involved in plant response to gravitropism (Li et al., 2007; Takeshi Yoshihara & Iino, 2007; T. Yoshihara, Spalding, & Iino, 2013) and TCP20s regulate plant basic cellular growth process (Guan et al., 2014; Johnson & Lenhard, 2011) (Tables S7, S9, S11). Further analysis on these genes is necessary. Similarly in the *W2* region, the *ATAUX2-11*-like gene M176753 (MD10G1193000) and the small auxin-up RNA (SAUR)-like genes M186167

(MD10G1202100) and M138076 (MD10G1204800) also appeared to be interesting candidates (**Table S2.8**, **Table S2.10**, **Table S2.11**). Several studies reported that *ATAUX2-11* is an auxin and gravitropic responsive transcription factor (Riechmann et al., 2000; Wyatt, Ainley, Nagao, Conner, & Key, 1993) and SAUR proteins are key regulators for plant growth and development (Ren & Gray, 2015), including plant branch angle (Bemer et al., 2017). Overall, the findings reported here represent an important step forward to a more comprehensive understanding of the weeping phenotype in *Malus*. However, due to the complex genome in *Malus*, a dedicated effort is required for revealing the identity of alleles *W* and *W2* and possibly *W3* and *W4*.

Conclusions

In an F_1 population developed in out-crossing woody species, there are at least three segregation types of DNA variants that are informative for genetic mapping using pooled genome sequencing analysis: <**l**m x mm> (Type-I) in the mutant (weeping) pool specific variants and <**l**m x **ll**> (Type-II) and <**h**k x **h**k> (Type-III) in the variants common to both mutant and wild-type pools. Type-I variants are commonly the first choice for mapping, and they are expected to include causal variants. Mapping using Type-I variants could be readily performed through MAFD mapping, however false positives may reduce the efficacy of this approach. Type-II and Type-III variants are important and more effective alternative for mapping, but causal variants are unlikely to be covered for dominant traits. AFDDD mapping is an effective approach to target Type-II and Type-III variants. Variant density appeared to be a better parameter than variant allele frequency in mapping. Both MAFD and AFDDD mappings can be applied for QTL discovery. There are four genomic regions of significant association with the weeping phenotype in *Malus*, including a major locus *Weeping* (*W*) on chromosome 13 and three others on chromosomes 10 (*W2*), 16 (*W3*) and 5 (*W4*). Confirmation of the mapping of *W* and

W2, investigation into their genetic interactions, and identification of expressed genes in the *W* and *W2* regions shed light on the genetic control of the weeping trait and its expressivity in *Malus*.

Acknowledgements

This work was financially supported by a grant award (IOS-1339211) from NSF-Plant Genome Research Program. A final version of the chapter is published in the Journal of Experimental Botany and can be found here 10.1093/jxb/erx490. Co-authors for the paper include Raksha Singh, Susan Brown, Chris Dardick and Kenong Xu.

REFERENCES

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., Matsumura, H., Yoshida, K., Mitsuoka, C., Tamiru, M., Innan, H., Cano, L., Kamoun, S., & Terauchi, R. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature biotechnology, 30*(2), 174-178.

Alston, F. H., Phillips, K. L., & Evans, K. M. (2000). A Malus Gene List. *Acta Hort. (ISHS), 538*, 561-570.

Bai, Y., Dougherty, L., & Xu, K. (2014). Towards an improved apple reference transcriptome using RNA-seq. *Mol Genet Genomics, 289*(3), 427-438.

Bemer, M., van Mourik, H., Muiño, J. M., Ferrándiz, C., Kaufmann, K., & Angenent, G. C. (2017). FRUITFULL controls SAUR10 expression and regulates Arabidopsis growth and architecture. *Journal of Experimental Botany*.

Brown, S. K. (1992). Genetics of Apple. In *Plant Breeding Reviews* (Vol. 9, pp. 333-366): John Wiley & Sons, Inc.

Brown, S. K., Maloney, K. E., Hemmat, M., & Aldwinckle, H. S. (2004, September 1-5, 2003). *Apple Breeding at Cornell: Genetic Studies of Fruit Quality, Scab Resistance and Plant Architecture.* Paper presented at the XI Eucarpia Symposium on Fruit Breeding and Genetics, Angers, France.

Carpita, N. C., & McCann, M. C. (2015). Characterizing visible and invisible cell wall mutant phenotypes. *Journal of Experimental Botany, 66*(14), 4145-4163.

Chaparro, J. X., Werner, D. J., O'Malley, D., & Sederoff, R. R. (1994). Targeted mapping and linkage analysis of morphological isozyme, and RAPD markers in peach. *Theoretical and Applied Genetics, 87*(7), 805-815.

Corredor-Moreno, P., Chalstrey, E., Lugo, C. A., & MacLean, D. (2015). IDENTIFICATION OF GENOMIC REGIONS CARRYING A CAUSAL MUTATION IN UNORDERED GENOMES. *bioRxiv*.

Costes, E., Lauri, P. E., & Regnard, J. L. (2006). Analyzing fruit tree architecture: implications for tree management and fruit production. *Horticultural Reviews, 32*, 1-61.

Dardick, C., Callahan, A., Horn, R., Ruiz, K. B., Zhebentyayeva, T., Hollender, C., Whitaker, M., Abbott, A., & Scorza, R. (2013). PpeTAC1 promotes the horizontal growth of branches in peach trees and is a member of a functionally conserved gene family found in diverse plants species. *The Plant Journal, 75*, 618–630.

Dirlewanger, E., & Bodo, C. (1994). Molecular genetic mapping of peach. *Euphytica, 77*(1), 101-103.

Duitama, J., Sánchez-Rodríguez, A., Goovaerts, A., Pulido-Tamayo, S., Hubmann, G., Foulquié-Moreno, M. R., Thevelein, J. M., Verstrepen, K. J., & Marchal, K. (2014). Improved linkage analysis of Quantitative Trait Loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast. *BMC Genomics, 15*(1), 207.

El-Sharkawy, I., Liang, D., & Xu, K. (2015). Transcriptome analysis of an apple (Malus × domestica) yellow fruit somatic mutation identifies a gene network module highly associated with anthocyanin and epigenetic regulation. *Journal of Experimental Botany, 66*, 7359–7376.

Fekih, R., Takagi, H., Tamiru, M., Abe, A., Natsume, S., Yaegashi, H., Sharma, S., Sharma, S., Kanzaki, H., Matsumura, H., Saitoh, H., Mitsuoka, C., Utsushi, H., Uemura, A., Kanzaki, E., Kosugi, S., Yoshida, K., Cano, L., Kamoun, S., & Terauchi, R. (2013). MutMap+:

Genetic Mapping and Mutant Identification without Crossing in Rice. *PLoS One, 8*(7), e68529.

Giovannoni, J. J., Wing, R. A., Ganal, M. W., & Tanksley, S. D. (1991). Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. *Nucleic Acids Res., 19*(23), 6553-6558.

Guan, P., Wang, R., Nacry, P., Breton, G., Kay, S. A., Pruneda-Paz, J. L., Davani, A., & Crawford, N. M. (2014). Nitrate foraging by Arabidopsis roots is mediated by the transcription factor TCP20 through the systemic signaling pathway. *Proceedings of the National Academy of Sciences, 111*(42), 15267-15272.

Hartwig, B., James, G. V., Konrad, K., Schneeberger, K., & Turck, F. (2012). Fast Isogenic Mapping-by-Sequencing of Ethyl Methanesulfonate-Induced Mutant Bulks. *Plant physiology, 160*(2), 591-600.

Hill, J. T., Demarest, B. L., Bisgrove, B. W., Gorsi, B., Su, Y.-C., & Yost, H. J. (2013). MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq. *Genome Research, 23*, 687–697.

Höfer, M., Flachowsky, H., Hanke, M.-V., Semënov, V., Šlâvas, A., Bandurko, I., Sorokin, A., & Alexanian, S. (2013). Assessment of phenotypic variation of Malus orientalis in the North Caucasus region. *Genetic Resources and Crop Evolution, 60*(4), 1463-1477.

Hollender, C. A., Hadiarto, T., Srinivasan, C., Scorza, R., & Dardick, C. (2016). A brachytic dwarfism trait (dw) in peach trees is caused by a nonsense mutation within the gibberellic acid receptor PpeGID1c. *New Phytologist, 210*(1), 227-239.

Hu, H., Wang, W., Zhu, Z., Zhu, J., Tan, D., Zhou, Z., Mao, C., & Chen, X. (2016). GIPS: A Software Guide to Sequencing-Based Direct Gene Cloning in Forward Genetics Studies. *Plant physiology, 170*(4), 1929-1934.

Huang, H. W., Mullikin, J. C., & Hansen, N. F. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics, 16*(1), 235.

Huo, H., Henry, I. M., Coppoolse, E. R., Verhoef-Post, M., Schut, J. W., de Rooij, H., Vogelaar, A., Joosen, R. V. L., Woudenberg, L., Comai, L., & Bradford, K. J. (2016). Rapid identification of lettuce seed germination mutants by bulked segregant analysis and whole genome sequencing. *The Plant Journal, Published online*.

Jensen, P., Fazio, G., Altman, N., Praul, C., & McNellis, T. (2014). Mapping in an apple (Malus x domestica) F1 segregating population based on physical clustering of differentially expressed genes. *BMC Genomics, 15*(1), 261.

Johnson, K., & Lenhard, M. (2011). Genetic control of plant organ growth. *New Phytologist, 191*(2), 319-333.

Just, B. J. (2001). *Molecular markers for weeping plant habit and powdery mildew (Podosphaera leucotricha) resistance from the ornamental crabapple 'Red Jade'.* (Master of Science). Cornell University,

Kaplow, I. M., MacIsaac, J. L., Mah, S. M., McEwen, L. M., Kobor, M. S., & Fraser, H. B. (2015). A pooling-based approach to mapping genetic variants associated with DNA methylation. *Genome Research, 25*(6), 907-917.

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics, 27*(24), 3435-3436.

Leshchiner, I., Alexa, K., Kelsey, P., Adzhubei, I., Austin-Tse, C. A., Cooney, J. D., Anderson, H., King, M. J., Stottmann, R. W., Garnaas, M. K., Ha, S. S., Drummond, I. A., Paw, B. H., North, T. E., Beier, D. R., Goessling, W., & Sunyaev, S. R. (2012). Mutation mapping and identification by whole-genome sequencing. *Genome Research, 22*(8), 1541-1548.

Lespinasse, J. M., & Delort, J. F. (1986). Apple tree management in vertical axis: appraisal after ten years of experiments. *Acta Horticulturae, 160*, 139-156.

Lespinasse, Y. (1992). *Breeding apple tree: aims and methods.* Paper presented at the Proceedings of the joint conference of the E.A.P.R breeding and varietal assessment section and the E.U.C.A.R.P.I.A. potato section, I.N.R.A., Ploudaniel (France).

Li, P. J., Wang, Y. H., Qian, Q., Fu, Z. M., Wang, M., Zeng, D. L., Li, B. H., Wang, X. J., & Li, J. Y. (2007). LAZY1 controls rice shoot gravitropism through regulating polar auxin transport. *Cell Research, 17*(5), 402-410.

Lindén, L., & Iwarsson, M. (2014). Identification of weeping crabapple cultivars by microsatellite DNA markers and morphological traits. *Scientia Horticulturae, 179*, 221-226.

Lindner, H., Raissig, M. T., Sailer, C., Shimosato-Asano, H., Bruggmann, R., & Grossniklaus, U. (2012). SNP-Ratio Mapping (SRM): Identifying Lethal Alleles and Mutations in Complex Genetic Backgrounds by Next-Generation Sequencing. *Genetics, 191*(4), 1381-U1476.

Lister, R., Gregory, B. D., & Ecker, J. R. (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology, 12*(2), 107-118.

Liu, S. Z., Yeh, C. T., Tang, H. M., Nettleton, D., & Schnable, P. S. (2012). Gene Mapping via Bulked Segregant RNA-Seq (BSR-Seq). *PLoS One, 7*(5), e36406.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics, 24*(3), 133-141.

Michelmore, R. W., Paran, I., & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA, 88*, 9828-9832.

Miller, A. C., Obholzer, N. D., Shah, A. N., Megason, S. G., & Moens, C. B. (2013). RNA-seq–based mapping and candidate identification of mutations from forward genetic screens. *Genome Research, 23*(4), 679-686.

Minevich, G., Park, D. S., Blankenberg, D., Poole, R. J., & Hobert, O. (2012). CloudMap: A Cloud-Based Pipeline for Analysis of Mutant Genome Sequences. *Genetics, 192*(4), 1249–1269.

Obholzer, N., Swinburne, I. A., Schwab, E., Nechiporuk, A. V., Nicolson, T., & Megason, S. G. (2012). Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing. *Development, 139*(22), 4280-4290.

Pereira-Lorenzo, S., Ramos-Cabrer, A. M., & Fischer, M. (2009). Breeding Apple (Malus× Domestica Borkh). In S. M. Jain & P. M. Priyadarshan (Eds.), *Breeding Plantation Tree Crops:Temperate Species* (pp. 1-49): Springer.

Petit, J., Bres, C., Mauxion, J.-P., Tai, F. W. J., Martin, L. B. B., Fich, E. A., Joubès, J., Rose, J. K. C., Domergue, F., & Rothan, C. (2016). The Glycerol-3-Phosphate Acyltransferase

GPAT6 from Tomato Plays a Central Role in Fruit Cutin Biosynthesis. *Plant physiology, 171*(2), 894-913.

Pulido-Tamayo, S., Duitama, J., & Marchal, K. (2016). EXPLoRA-web: linkage analysis of quantitative trait loci using bulk segregant analysis. *Nucleic Acids Research*, Published online.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics, 13*(1), 341.

Ren, H., & Gray, William M. (2015). SAUR Proteins as Effectors of Hormonal and Environmental Signals in Plant Growth. *Molecular Plant, 8*(8), 1153-1164.

Ribeiro, A., Golicz, A., Hackett, C. A., Milne, I., Stephen, G., Marshall, D., Flavell, A. J., & Bayer, M. (2015). An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. *BMC Bioinformatics, 16*(1), 1-16.

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C. Z., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O. J., Samaha, R. R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J. Z., Ghandehari, D., Sherman, B. K., & Yu, G. L. (2000). Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. *Science (Washington D C), 290*(5499), 2105-2110.

Roberts, D. J., Werner, D. J., Wadl, P. A., & Trigiano, R. N. (2015). Inheritance and allelism of morphological traits in eastern redbud (Cercis canadensis L.). *Horticulture Research, 2*, 15049.

Sampson, D. R., & Cambron, D. F. (1965). Inheritance of bronze foliage, extra petals and pendulous habit in ornamental crabapples. *Proceedings. American Society for Horticultural Science, 86*, 717-722.

Sandal, N., Jin, H., Rodriguez-Navarro, D. N., Temprano, F., Cvitanich, C., Brachmann, A., Sato, S., Kawaguchi, M., Tabata, S., Parniske, M., Ruiz-Sainz, J. E., Andersen, S. U., & Stougaard, J. (2012). A Set of Lotus japonicus Gifu x Lotus burttii Recombinant Inbred Lines Facilitates Map-based Cloning and QTL Mapping. *DNA Research, 19*(4), 317-323.

Sarin, S., Prabhu, S., O'Meara, M. M., Pe'er, I., & Hobert, O. (2008). Caenorhabditis elegans mutant allele identification by whole-genome sequencing. *Nat Meth, 5*(10), 865-867.

Sato, A., Sasaki, S., Matsuzaki, J., & Yamamoto, K. T. (2014). Light-dependent gravitropism and negative phototropism of inflorescence stems in a dominant Aux/IAA mutant of Arabidopsis thaliana, axr2. *Journal of Plant Research, 127*(5), 627-639.

Sato, A., Sasaki, S., Matsuzaki, J., & Yamamoto, K. T. (2015). Negative phototropism is seen in Arabidopsis inflorescences when auxin signaling is reduced to a minimal level by an Aux/IAA dominant mutation, axr2. *Plant Signaling & Behavior, 10*(3), e990838.

Schierenbeck, L., Ries, D., Rogge, K., Grewe, S., Weisshaar, B., & Kruse, O. (2015). Fast forward genetics to identify mutations causing a high light tolerant phenotype in Chlamydomonas reinhardtii by whole-genome-sequencing. *BMC Genomics, 16*(1), 1-15.

Schneeberger, K. (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat Rev Genet, 15*(10), 662-676.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jorgensen, J. E., Weigel, D., & Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods, 6*(8), 550-551.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat Meth, 5*(1), 16-18.

Shaw, S. H., Carrasquillo, M. M., Kashuk, C., Puffenberger, E. G., & Chakravarti, A. (1998). Allele Frequency Distributions in Pooled DNA Samples: Applications to Mapping Complex Disease Genes. *Genome Research, 8*(2), 111-123.

Sun, H., & Schneeberger, K. (2015). SHOREmap v3.0: Fast and Accurate Identification of Causal Mutations from Forward Genetic Screens. In J. M. Alonso & A. N. Stepanova (Eds.), *Plant Functional Genomics* (Vol. 1284, pp. 381-395): Springer New York.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S., & Terauchi, R. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal, 74*(1), 174-183.

Trick, M., Adamski, N. M., Mugford, S. G., Jiang, C. C., Febrer, M., & Uauy, C. (2012). Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biology, 12*, 14.

Urbanowicz, B. R., Peña, M. J., Ratnaparkhe, S., Avci, U., Backe, J., Steet, H. F., Foston, M., Li, H., O'Neill, M. A., Ragauskas, A. J., Darvill, A. G., Wyman, C., Gilbert, H. J., & York, W. S. (2012). 4-O-methylation of glucuronic acid in Arabidopsis glucuronoxylan is catalyzed by a domain of unknown function family 579 protein. *Proceedings of the National Academy of Sciences, 109*(35), 14253-14258.

Van Leeuwen, T., Demaeght, P., Osborne, E. J., Dermauw, W., Gohlke, S., Nauen, R., Grbic, M., Tirry, L., Merzendorfer, H., & Clark, R. M. (2012). Population bulk segregant mapping uncovers resistance mutations and the mode of action of a chitin synthesis

inhibitor in arthropods. *Proceedings of the National Academy of Sciences of the United States of America, 109*(12), 4407-4412.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L. M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J. T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V. T., King, S. T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M. M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A. C., Bus, V., Chagne, D., Crowhurst, R. N., Gleave, A. P., Lavezzo, E., Fawcett, J. A., Proost, S., Rouze, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R. P., Durel, C.-E., Gutin, A., Bumgarner, R. E., Gardiner, S. E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F., & Viola, R. (2010). The genome of the domesticated apple (Malus x domestica Borkh.). *Nat Genet*, *42*, 833-839.

Wang, A., Aldwinckle, H., Forsline, P., Main, D., Fazio, G., Brown, S., & Xu, K. (2012). EST contig-based SSR linkage maps for Malus × domestica cv Royal Gala and an apple scab resistant accession of M. sieversii, the progenitor species of domestic apple. *Molecular Breeding*, *29*, 379-397.

Wyatt, R. E., Ainley, W. M., Nagao, R. T., Conner, T. W., & Key, J. L. (1993). Expression of the Arabidopsis AtAux2-11 auxin-responsive gene in transgenic plants. *Plant Molecular Biology*, *22*(5), 731-749.

Xu, K., & Mackill, D. J. (1996). A major locus for submergence tolerance mapped on rice chromosome 9. *Mol. Breeding, 2*(3), 219-224.

Xu, K., Wang, A., & Brown, S. (2012). Genetic characterization of the Ma locus with pH and titratable acidity in apple. *Molecular Breeding, 30*(2), 899–912.

Xu, K., Xu, X., Ronald, P. C., & Mackill, D. J. (2000). A high-resolution linkage map in the vicinity of the rice submergence tolerance locus *Sub1*. *Mol. Gen. Genet., 263*(4), 681-689.

Yang, J., Jiang, H., Yeh, C.-T., Yu, J., Jeddeloh, J. A., Nettleton, D., & Schnable, P. S. (2015). Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. *The Plant Journal, 84*(3), 587-596.

Yoshihara, T., & Iino, M. (2007). Identification of the Gravitropism-Related Rice Gene LAZY1 and Elucidation of LAZY1-Dependent and -Independent Gravity Signaling Pathways. *Plant and Cell Physiology, 48*(5), 678-688.

Yoshihara, T., Spalding, E. P., & Iino, M. (2013). AtLAZY1 is a signaling component required for gravitropism of the Arabidopsis thaliana inflorescence. *Plant J, 74*(2), 267-279.

Yu, H., Moss, B. L., Jang, S. S., Prigge, M., Klavins, E., Nemhauser, J. L., & Estelle, M. (2013). Mutations in the TIR1 Auxin Receptor That Increase Affinity for Auxin/Indole-3-Acetic Acid Proteins Result in Auxin Hypersensitivity. *Plant physiology, 162*(1), 295-303.

Zhang, J., Zhang, Q., Cheng, T., Yang, W., Pan, H., Zhong, J., Huang, L., & Liu, E. (2015). High-density genetic map construction and identification of a locus controlling weeping trait in an ornamental woody plant (Prunus mume Sieb. et Zucc). *DNA Research, 22*(3), 183-191.

Zuryn, S., Le Gras, S., Jamet, K., & Jarriault, S. (2010). A Strategy for Direct Mapping and Identification of Mutations by Whole-Genome Sequencing. *Genetics, 186*(1), 427-430.

Supplementary Figures

Figure S2.1. A typical weeping and a standard F_1 progeny from population 'Cheal's Weeping' × 'Evereste' after being budded for 1.5 years on apple rootstock B118.

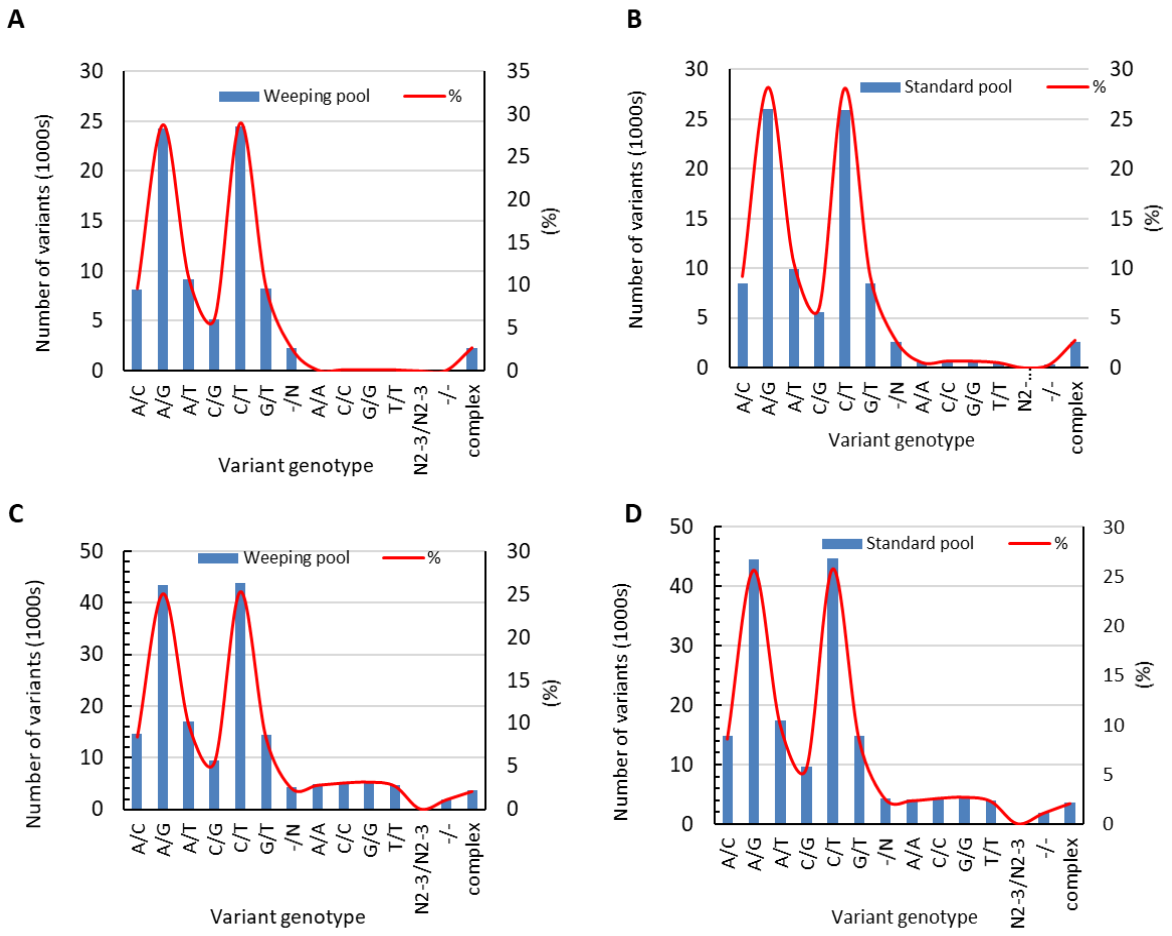


Figure S2.2. Genotype frequency of DNA variants specific to the weeping (**A**) and standard (**B**) pools and common to both weeping (**C**) and standard (**D**) pools from population 'Cheal's Weeping' x 'Evereste'. 'N2-3' and '-'stand for 2- or 3-nucleotide variants and InDels, respectively. In the pool specific variants, heterozygous variants comprising seven groups A/C, A/G, A/T, C/G, C/T, G/T, and -/N (heterozygous InDel) were predominant, accounting for 96.6% in weeping and 94.4% in standard pools (**A**, **B**). Variant genotypes A/G and C/T were most common and each explained 28.2-28.9%. In the variants common to both pools, a similar trend was observed. However, heterozygous variants were lower, denoting 84.8% in weeping and 86.8% in standard while homozygous variants, mainly A/A, C/C, G/G and T/T, accounted for 13.1% and 11.1% in weeping and standard pools, respectively (**C**, **D**).

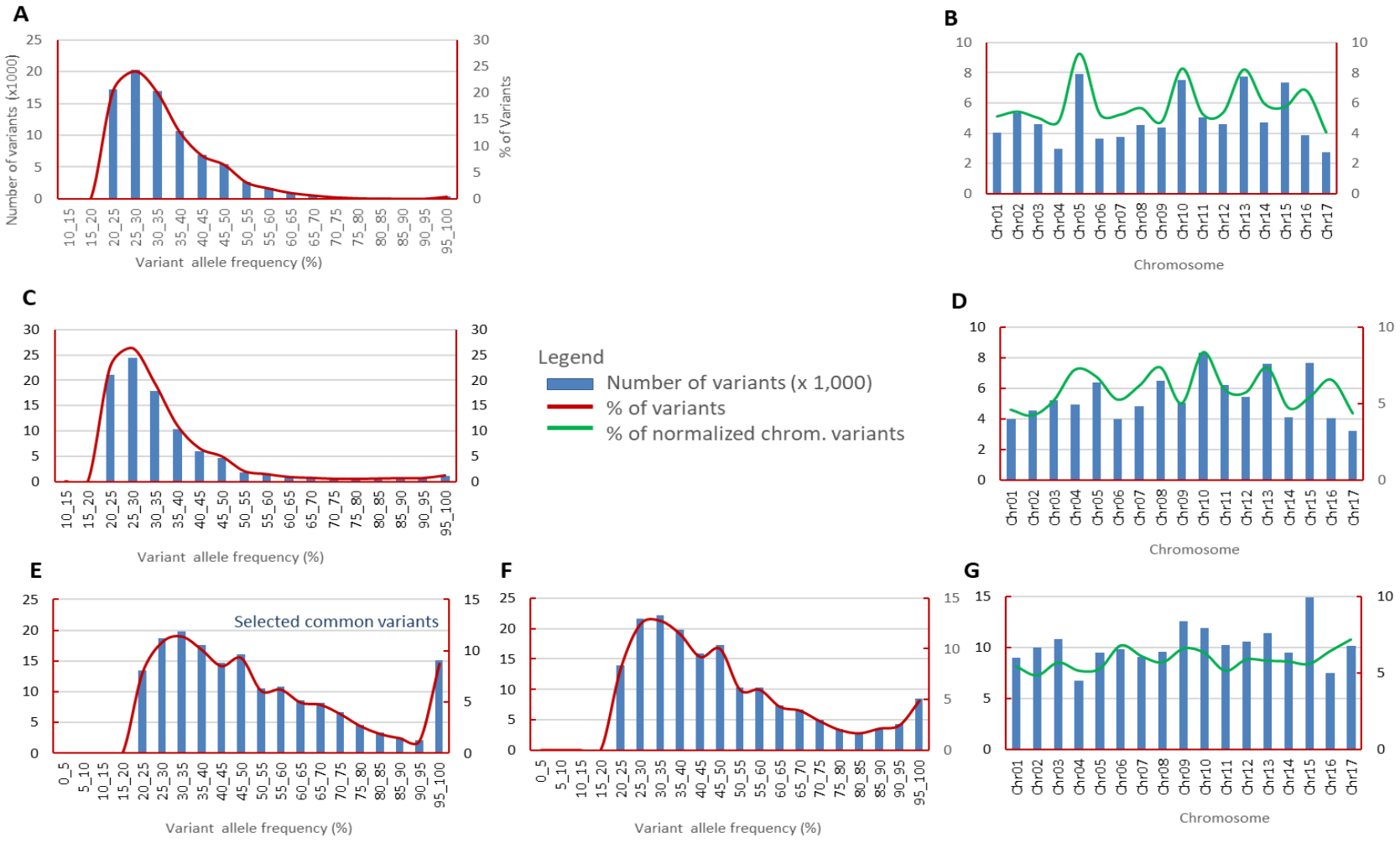


Figure S2.3. Distribution of pool specific and common variants**.** (**A**-**B**) Distribution of weeping pool specific variants (84,562) by allele frequency (**A**) and chromosome (**B**). (**C**-**D**) Distribution of standard pool specific variants (92,148) by allele frequency (**C**) and chromosome (**D**). (**E**-**G**) Distribution of the common variants (173,169) between the weeping (**E**) and standard (**F**) pools by allele frequency, and by chromosome (**G**). On all charts, the primary vertical axis is for number of variants in 1000s, and the secondary

vertical axis is for percent of variants. The percentage of normalized chromosomal variants (shown by the green curves in **B**, **D** and **G**) was calculated by factoring in chromosome physical size so that the data could be directly compared among chromosomes. Two points could be made by examining the variant allele frequency distribution and variant chromosomal distribution: 1) The variants specific to the weeping and standard pools had a highly similar allele frequency distribution with most ranging from 20% to 50% (**A**, **C**), suggesting that the selection in establishment of the pools did not lead to much difference between the weeping and standard pools in allele frequency distribution of pool specific variants. The similarities were also shown in the normalized chromosomal distributions with chromosomes 10, 13 and 16 having relatively more variants in both pools (**B**, **D**). However, there were more variants on chromosome 5, less on chromosomes 4 and 8 in the weeping pool than in the standard pool. 2) The variants common to both pools had much wider spread of allele frequency, ranging from 20% to 100% (**E**, **F**). Although the variants also showed similarities in allele frequency distribution between the two pools, there were markedly more variants (homozygous) of allele frequency 95-100% in the weeping pool (15,143 or 8.7%) than in standard pool (8,446, or 4.9%) (**E**, **F**). The normalized chromosomal distributions were largely similar among chromosomes in variants common to both pools, varying narrowly from 4.9% (chromosome 2) to 7.2% (chromosome 17), close to the average distribution of 5.9%, suggesting the number of variants identified from each chromosome was roughly even (**G**).

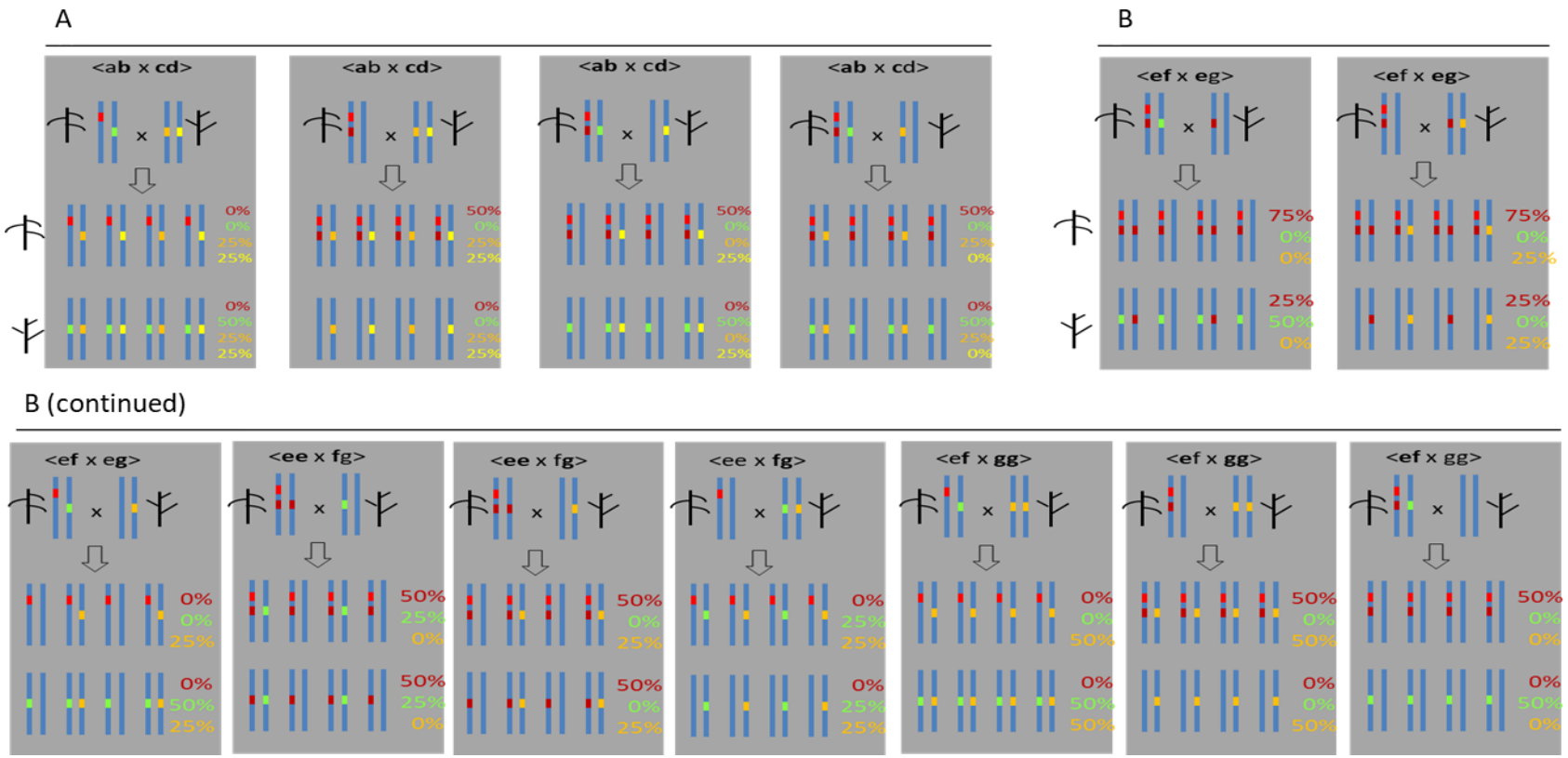


Figure S2.4. Schematics for possible segregation types inferred for variant genotype group G5 'Complex'. (A) <ab x cd> derived segregation types involving four DNA bases (or three DNA variants). (B) <ef x eg> derived segregation types involving three DNA bases (or two DNA variants). Each segregation type is illustrated in a grey filled rectangular, which includes the two parents at the top, four representative weeping progeny in the mid, and four standard progeny at the bottom. The long vertical lines in blue stand for the chromosomal segment harboring *W*. The short vertical lines in red and in other colors (orange, purple, and green) represent allele *W* and DNA variants in relation to the reference genome, respectively. The tree-like drawings with up- and down-ward 'branches' indicate standard and weeping tree phenotypes, respectively. The expected allele frequency of DNA variants in the weeping and standard pools is given and color coded accordingly. In each segregation type denotation, the allele at the first position is designated to be linked to weeping phenotype in the seed parent 'Cheal's Weeping' (e.g. letter 'a' in <**a**b **x cd**>), and those in bold are DNA variants in relation to the apple reference genome (e.g. letters 'a, c and d' in <**a**b **x cd**>).

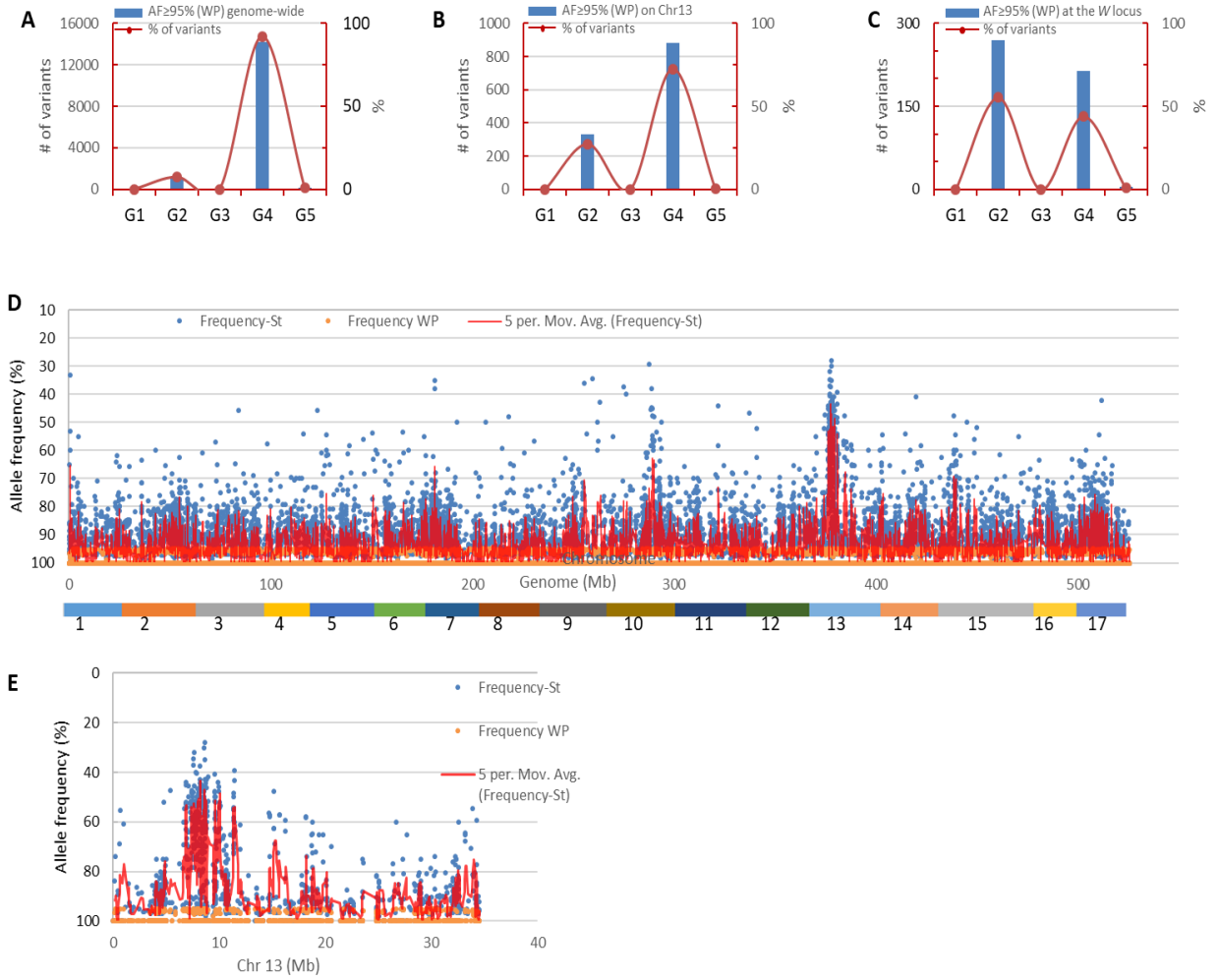


Figure S2.5. Evaluating the role of segregation type <**l**m x **ll**> (Type-II variants) in AFDDD mapping using common variants (15,425) selected of allele frequency ≥95% in the weeping pool (WP). (**A**-**C**) Number and frequency (%) of such selected variants observed in the five genotype groups at the genome scale (**A**), on chromosome 13 (**B**) and in the W region (**C**). G1: heterozygous in weeping / heterozygous in standard (He-W/He-S); G2: Ho-W/He-S; G3: He-W/Ho-S; G4: Ho-W/Ho-S; and G5: 'Complex'. (**D**-**E**) Allele frequency distribution of the variants in the weeping and stand pools in the genome (**D**) and on chromosome 13 (**E**). The color bar represents the assembled reference genome of 17 chromosomes as numbered. Notes of analysis: The composition of the 15,425 variants was dominated by homozygous variants in G4 of segregation type <**qq** x **qq**>, which accounted for 92.0% (14,185) (A). The second category was the supposed-to-be targeted genotype group G2 by the selection, weighed 7.3% (1,121). Although the weight of this variant genotype group was low, it represented an increase of 135.0% when compared with 3.1% prior to the selection (see **Figure. 2.7D**). Plotting the 15,425 variants from both pools against the genome uncovered that the distribution of variants on chromosome 13 was most distinctive from those on any other chromosomes (**D**). There were 1,219 variants on chromosome 13 (**B**, **E**) and 487 in the 7-Mb region of *W* (**C**). Among these two sets of variants, the frequency of the targeted G2 genotype group were drastically increased to 27.2% (332/1,219) and 54.0% (269/487), respectively (**B**, **C**). Examining the 269 variants in the

standard pool revealed that there were 79 of allele frequency close to the expected 50%, ranging from 45% to 55% (**E**). Since the allele frequency range from 45% to 55% differs at least by 40 percentage points from the allele frequency threshold ≥95% used for selection in the weeping pool and there were only 682 variants of AFDD≥40 percentage points in the G2 genotype group genome-wide, the finding of these 79 variants in the *W* region (**E**) represented a highly significant increase (z=14.1, p=0) of such expected allele frequency profile for Type-II variants in the *W* region of 7-Mb, providing convincing evidence that Type-II variants are an essential part in AFDDD mapping.

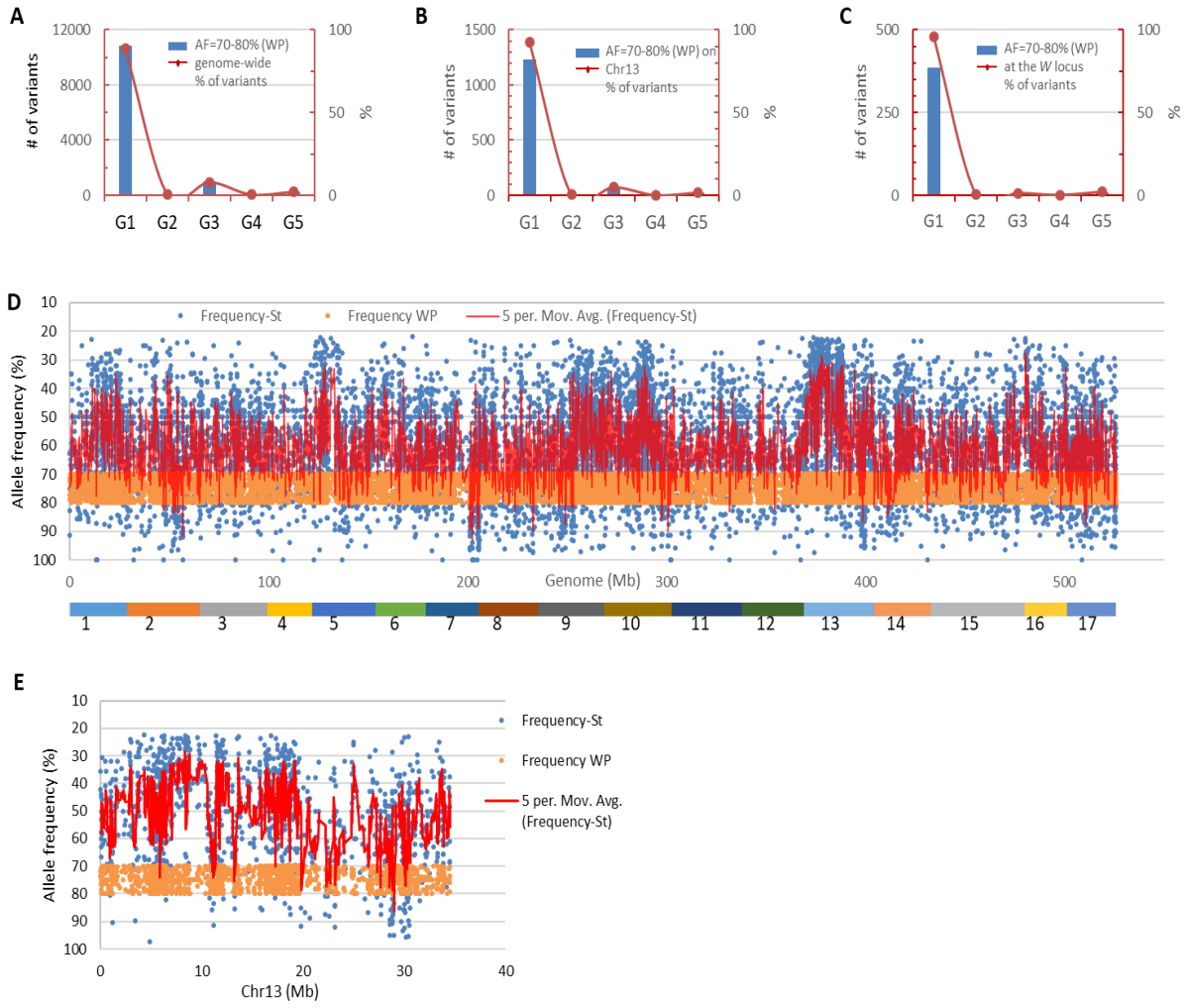


Figure S2.6. Evaluating the role of segregation type <**h**k x **h**k> (Type-III variants) in AFDDD mapping using common variants (12,219) selected of allele frequency ranging from 70% to 80% in the weeping pool (WP). (**A**-**C**) Number and frequency (%) of such selected variants observed in the five genotype groups at the genome scale (**A**), on chromosome 13 (**B**) and in the W region (**C**). G1: heterozygous in weeping / heterozygous in standard (He-W/He-S); G2: Ho-W/He-S; G3: He-W/Ho-S; G4: Ho-W/Ho-S; and G5: 'Complex'. (**D**-**E**) Allele frequency distribution of the variants in the weeping and stand pools in the genome (D) and on chromosome 13 (**E**). The color bar represents the assembled reference genome of 17 chromosomes as numbered. Notes of analysis: Among the 12,219 variants, 89.0% (10,869/12,219) fell in the genotype group G1 (**A**), suggesting that the selection in the weeping pool not only included Type-III variants, but also led to a 5.5-percentage-point increase from 83.5% in the 173,169 variants (see **Figure. 2.7 D**). Plotting the 12,219 variants from both pools against the reference genome highlighted that there was a clustered group of 58 variants of allele frequency from 20% to 30% (close to the expected 25%) in the 7-Mb region of *W* on chromosome 13 (**D**, **E**). Since the allele frequency threshold ≤30% differs at least by 40 percentage points from the 70%-80% range used for selection in the weeping pool and there were only 908 variants of the G1 genotype group with AFDD≥40 percentage points genome-wide, these 58 variants represented a significant increase (z=8.1, p=0) in the *W* region. In addition, the frequency of the variants of in G1 was 92.6% (1233/1311) on

chromosome 13 (B) and 96.0% (385/401) in the 7-Mb region of *W* (**C**). This was much higher than the original 83.5% (see **Figure. 2.7 D**), contrasting the overall decrease trend in AFDDD mapping (see **Figure. 2.7 A**-**C**). These observations indicated that Type-III variants were selected preferentially in the *W* region due to pooling, i.e. Type-III variants also played a critical role in AFDDD mapping.

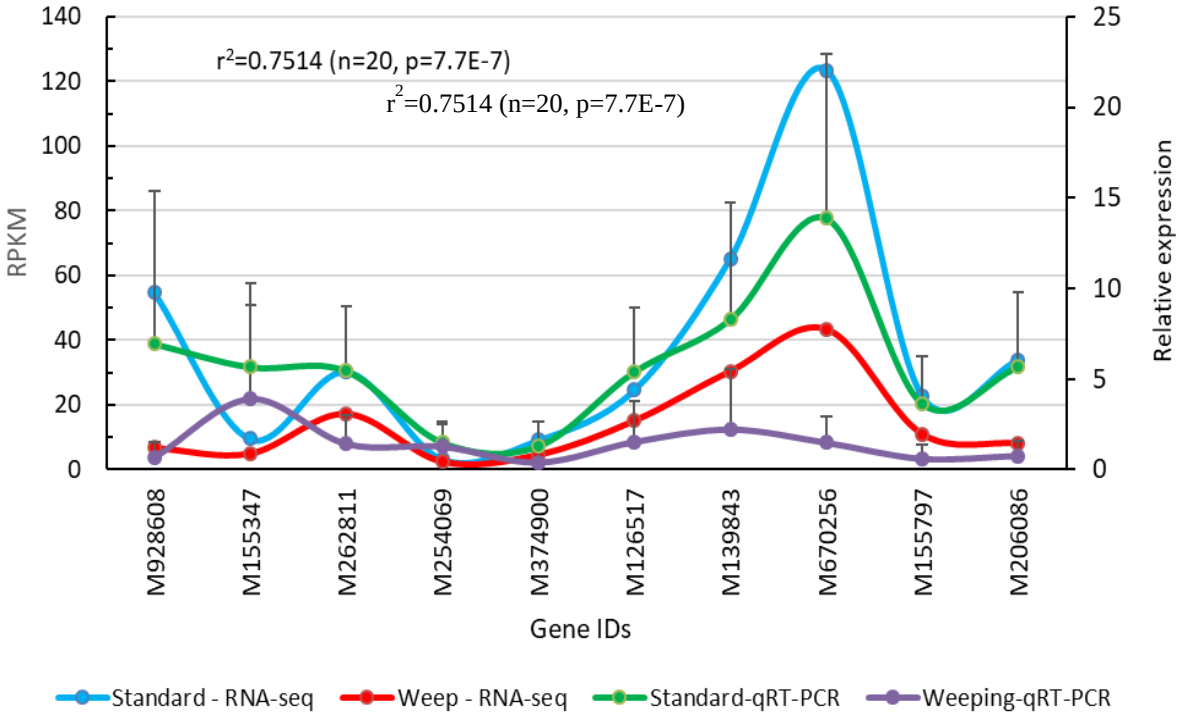


Figure S2.7. qRT-PCR validation of gene expression in RNA-seq analysis. The qRT-PCR assays were performed on ten genes in the *W* region. The expression levels of the ten genes in RPKMs in RNA-seq analysis are highly significantly correlated with their relative expression in qRT-PCR assays (r^2=0.7514 (n=20, p=7.7E-7). RPKM: Reads Per Kilobase of transcript per Million mapped reads. The prefix 'MDP0000' in Apple gene Ids (e.g. MDP0000123456) is abbreviated to 'M' (e.g. M123456).

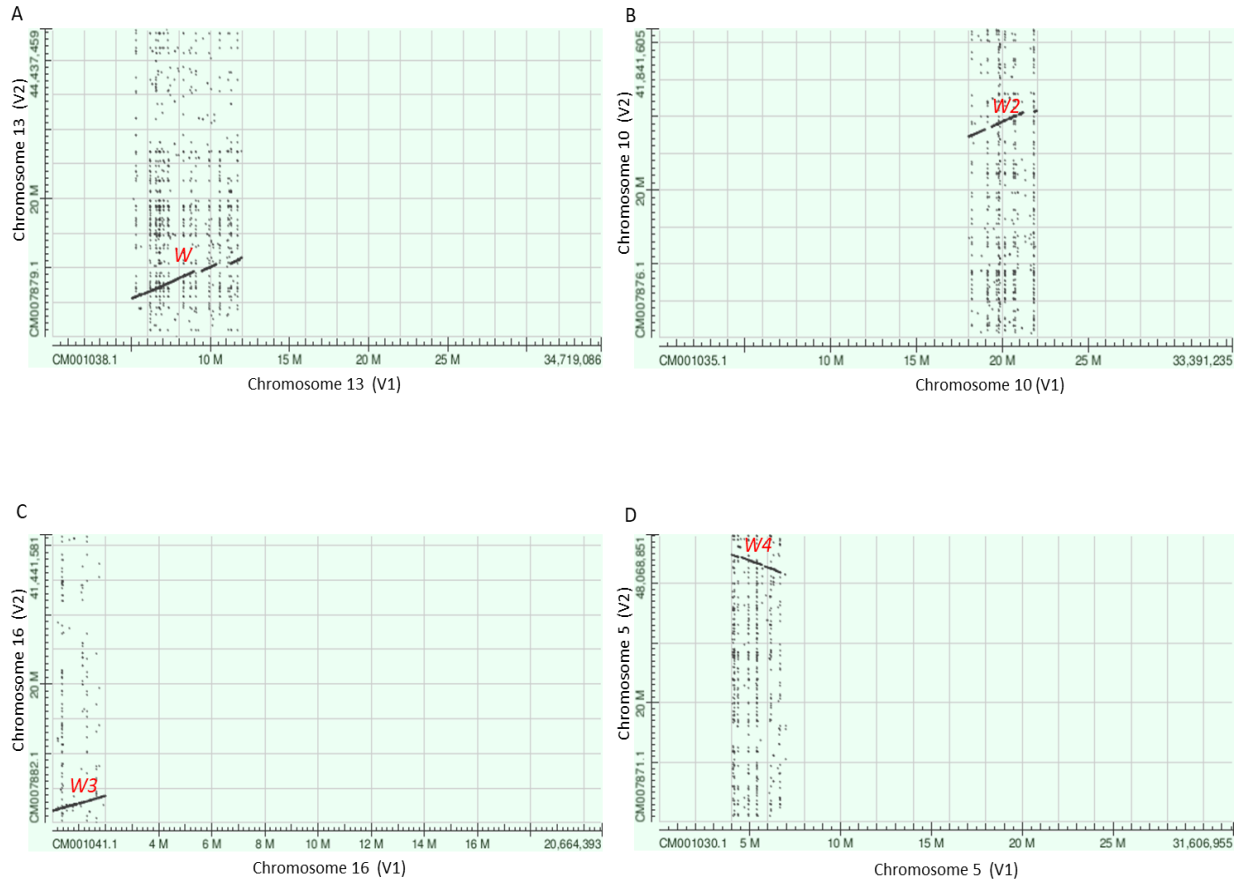


Figure S2.8. BLAST-based dot matrix analysis of the genomic regions associated with the weeping trait. The X- and Y-axis and the numbers represent the bases from chromosomes 13 (**A**), 10 (**B**), 16 (**C**) and 5 (**D**) in the first version (V1, Velasco et al. 2010) and the new version (V2, Daccord et al. 2017) of the apple reference genome, respectively. The *W, W2*, *W3* and *W4* regions in the V1 and V2 are indicated by dotted lines, which represent genomic regions of high levels of sequence identity between V1 and V2.

Supplementary Tables

Table S2.1. Reads Mapping Summary

Pools	Reads category	Count	Percentage of reads	Average length	Number of bases	Percentage of bases
Weeping pool	Mapped reads	57,639,266	43.2%	151	8,703,529,166	43.2%
	Not mapped reads	75,677,546	56.8%	151	11,427,309,446	56.8%
	Reads in pairs	29,551,548	22.2%	532.81	4,462,283,748	22.2%
	Broken paired reads	28,087,718	21.1%	151	4,241,245,418	21.1%
	Total reads	133,316,812	100.0%	151	20,130,838,612	100.0%
Standard pool	Mapped reads	66,798,158	44.7%	151	10,086,521,858	44.7%
	Not mapped reads	82,641,060	55.3%	151	12,478,800,060	55.3%
	Reads in pairs	35,549,144	23.8%	514.52	5,367,920,744	23.8%
	Broken paired reads	31,249,014	20.9%	151	4,718,601,114	20.9%
	Total reads	149,439,218	100.0%	151	22,565,321,918	100.0%

Table S2.2. Variant filtering process

Filters	Weeping	Standard	Common
Total variants: Detection of variants using CLC Genomics Workbench variant detection tools	2,700,059	2,946,289	
Variants non-allelic to reference: Remove variants of reference alleles (including M, R, W, S, Y, K), being hyper allelic or homopolymer	1,306,887 (SNV: 88.5%)	1,380,503 (SNV: 87.8%)	
Pool specific/common variants: Compare the variants between the weeping and standard pools	498,386	573,589	799,089
Putative variants for mapping: Coverage ≥19 (weeping) or ≥20 (standard); Forward/reverse balance: 0.25-0.5. Number of reads with unique start positions: ≥5	84,562	92,148	173,169

Table S2.3. Primer sequences and their genome physical locations

Primer Name	Marker name/purpose	Sequence (5' to 3')	Targeted positions on chromosomes
Ch13-7641F	Ch13-7641-119bp	TTCGCCTAGTTTGGTCCGTCA	7.641th Mb on chr13
Ch13-7641R		GGGTCCCTGAGAGTCCAGTGC	
Ch13-8181F	Ch13-8181-125bp	TCTTCGAACACACCCGCAAA	8.181th Mb on chr13
Ch13-8181R		GGTTAATGCGCACCGGGTTA	
Ch13-8547F	Ch13-8547-168bp	CCGACCCCAAATGCGTTTAT	8.547th Mb on chr13
Ch13-8547R		GTCCCTGAATTATTCACCACAA	
Ch13-9530F	Ch13-9530-175bp	TTTCTCCGTCCATGTCCTTGA	9.530th Mb on chr13
Ch13-9530R		CCATGGTTGTACTGCGTTTTC	
Ch10-19768F	Ch10-19768-190bp	TTTGGGTCTCCAAATGCATAG	19.768th Mb on chr10
Ch10-19768R		GCTATGTTCAGCTCGTACCG	
Ch10-20017F	Ch10-20017-220bp	GGAATGTTTTGAGTGGTGTCA	20.017th Mb on chr10
Ch10-20017R		GGGAGGGGTGAAGATTCAGT	
Ch13_7923F	SNP validation	CCTTCCGTCTATACCCAGCA	7923174, **7923460**, 7923530 on chr13
Ch13_7923R		TTGAACTCGGATGCAAATCA	
Ch13_8209F	SNP validation	TCGATGAAATTTGCTGTGAAA	**8209678**, **8210175 on chr13**
Ch13_8209R		TTCTCCAAAACTGAGGCAAA	
Ch13_8374F	SNP validation	CTACAGGGAAACCGCTCAAG	8374569, 8375098, 8375311, 8375431, 8375828, 8375948 on chr13
Ch13_8374R		AGCAAGCAAACCATCCTTGT	
Ch13_8758F	SNP validation	GAGCACGGGTTATGGAAGAA	8758311, 8759078, 8759167 on chr13
Ch13_8758R		GCACTGCACGTAATCAAACG	

[1]Positions in bold also were shown in Figure. 8

Table S2.4. qRT-PCR primer sequences and their targeted gene IDs

Gene IDs	Primer sequences (F/R)	Prod size
MDP0000928608	CACGTGTTCCTTCACGATGT / GCTAACGGGCCAAATATCCT	300
MDP0000155347	GGCGAATCTGTACCAGGAAA / ACGCATAGTTCAACCGGAAA	296
MDP0000262811	CATTCAACAGGCCAACAATG / AAGAAGAAGATGGCCACAGC	300
MDP0000374900	GGACATCCCTTGGTGAGCTA / TCTTGGTTTGGTTCCTTTCG	303
MDP0000126517	GGGAAATGGGTTGTTTTCCT / TTCCCCAATGAAGGACTCTG	299
MDP0000254069	GGGAACAGATCGAACAGAGC / TTTTTGCCTCCCCTTTTCTT	299
MDP0000139843	CTCCAAATCCCAATTCCAGA / GGTGCCGTTGTAGAAAATCG	301
MDP0000670256	ATGGCTTCGAGTTCTGCAAC / CCAAGCTCATTGATCCTTTTC	241
MDP0000155797	CCGGTTGCTATCTGGTTTGT / TCAAGGCCATCTTCTCGTCT	301
MDP0000206086	TCCATATGCTCCACCACAGA / GGATGCAGCCAAATACCACT	305

Table S2.5. Genotype groups of variants common to both pools and variant segregation types inferred

Genotype group	Variant genotypes observed[a]			Inferred[b]								Notes
	W pool	S pool	# of variants (freq.)	Segregation types (genotype of parents)[c]	W pool genotype		W pool mean AF (%)	S pool genotype		S pool mean AF (%)	AFD between W and S pools (%)	
G1	He-W	He-S	144,558 (83.5%)	<**h**k x **h**k>	hh	hk	75	hk	kk	25	50	Informative for *W*
				<h**k** x h**k**>	hh	hk	25	hk	kk	75	-50	Informative for *w*
				<**nn** x **n**p>	nn	np	75	nn	np	75	0	Not informative
				<nn x n**p**>	nn	np	25	nn	np	25	0	Not informative.
				<**hh** x kk>	hk	hk	50	hk	hk	50	0	Not informative
				<hh x **kk**>	hk	hk	50	hk	hk	50	0	Not informative
G2	Ho-W	He-S	5,353 (3.1%)	<**l**m x **ll**>	ll	ll	100	ml	ml	50	50	Informative for *W*
				<l**m** x ll>	ll	ll	0	ml	ml	50	-50	Informative for *w* using S-pool specific variants
G3	He-W	Ho-S	2,104 (1.2%)	<l**m** x **mm**>	lm	lm	50	mm	mm	100	-50	Informative for *w*
				<**l**m x mm>	lm	lm	50	mm	mm	0	50	Informative for *W* using W-pool specific variants
G4	Ho-W	Ho-S	16,963 (9.8%)	<**qq** x **qq**>	qq	qq	100	qq	qq	100	0	Not informative
G5	complex	complex	4,191 (2.4%)	<e**f** x e**g**>	complex			complex				Complex
				<**ab** x **c**d>	complex			complex				Complex

[a] Homozygous (Ho): variant allele frequency (AF)>80%; Heterozygous (He): 80%> AF>15%; [b] for variants in the *W* region. [c] The alleles in each first position are designated to be linked to weeping phenotype in seed parent 'Cheal's Weeping' and those in bold are a

polymorphic variant in relation to the apple reference genome, which are present in both parents. If the variants are from the seed parent 'Cheal's Weeping', they can be linked to the weeping phenotype in either coupling phase or repulsion phase. W: weeping; S or w: standard; AFD: allele frequency difference. Notes for inferring segregation types for variants in genotype groups G1-G3: G1 (He-W/He-S) was inferred to be caused by six possible variant segregation types, including <**h**k x **h**k>, <h**k x** h**k**>, <**nn** x **n**p>, <nn x n**p**>, <**hh** x kk> and <hh x **kk**>. Of these, the first segregation type <**h**k x **h**k> is the only one that would be informative for mapping allele *W* as it would confer an average '**h**' allele frequencies 75% and 25% in the weeping and standard pools, respectively, allowing a directional (positive) 50-percentage-point difference in allele frequency relative to that in the weeping pool (see also Figure. 3A). The second segregation type <h**k x** h**k**> is informative for the standard phenotype (allele *w*) for a similar but negative 50-percentage-point difference in allele frequency between the weeping and standard pools (see also Figure. 3A). Since this segregation type is not useful for mapping allele *W*, variants of segregation type <h**k x** h**k**> were excluded from further analysis. The latter four were non-informative for mapping as seed parent 'Cheal's Weeping' is homozygous and equal levels of allele frequencies (25%, 50% or 75%) are expected between the two pools, but they were likely the major source for heterozygous variants in the pools based on their expected allele frequencies and the distribution of what actually was observed (see also Figure. S3A, C, E, F). G2 (Ho-W/He-S) was inferred with two possible segregation types <**l**m x **ll**> and <l**m** x ll>. The former, <**l**m x **ll**> is informative for mapping alleles *W* as the allele frequency of '**l**' is expected to be 100% in weeping pool, and 50% in the standard pool, also allowing a positive 50-percentage difference in allele frequency (see also Figure. 3B). The latter, <l**m** x ll> is informative for allele *w* as it would give the allele frequency of '**m**' 0% in the weeping pool, and 50% in the standard pool, leading to a negative 50-percentage difference in allele frequency. It could be used for variants specific to the standard pool, but similar to <h**k x** h**k**>, variants of segregation <l**m** x ll> may not be helpful for mapping allele *W*. G3 (He-W/Ho-S), which is opposite to G2, was similarly inferred with two possible segregation types <**l**m x mm> and <l**m** x **mm**>. In this case, the former would confer a 50% average '**l**' allele frequency in weeping pool specific variants (0% in the standard pool), making segregation type <**l**m x mm> informative for mapping allele *W* (see also Figure. 3C). The latter <l**m** x **mm**> would produce a negative 50-pecentage-point difference in '**m**' allele frequency (50% in weeping, 100% in standard). Again, similar to <h**k x** h**k**>, variants of segregation type <l**m** x **mm**> were disregarded for mapping allele *W*.

Table S2.6. Summary of RNA-seq reads mapping

RNA-seq sample	Reads category	Apple Reference Genome V1[a]		Apple Reference Genome V2[b]	
		No. of Reads	% of clean reads	No. of Reads	% of clean reads
Standard	Clean reads	52,423,649	100	52,423,649	100
	Mapped reads	39,552,212	75.5	39,606,583	75.6
	mapped uniquely	31,665,782	60.4	37,644,378	71.8
	mapped non-uniquely	7,886,430	15	1,962,205	3.7
	Unmapped reads	12,871,437	24.6	12,817,066	24.5
	Low quality reads	3,802,514		3,802,514	
	rRNA reads	2,932,563		2,932,563	
	Total	59,158,726		59,158,726	
Weeping	Clean reads	39,916,890	100	39,916,890	100
	Mapped reads	30,269,137	75.8	30,558,185	76.6
	mapped uniquely	24,305,524	60.9	28,973,044	72.6
	mapped non-uniquely	5,963,613	14.9	1,585,141	4.0
	Unmapped reads	9,647,753	24.2	9,358,705	23.5
	Low quality reads	2,031,399		2,031,399	
	rRNA reads	1,260,009		1,260,009	
	Total	43,208,298		43,208,298	

[a]Apple Reference Genome V1: Velasco et al. 2010 and Bai et al 2014; [b]Apple Reference Genome V2: Daccord et al 2017

Table S2.7. List of expressed genes in the *W* region

Table is provided as a supplementary spreadsheet

Tble S2.8. List of expressed genes in the *W2* region

Table is provided as a supplementary spreadsheet.

Table S2.9. List of expressed genes in the *W* region according to the new reference genome (Daccord et al. 2017)

Table is provided as a supplementary spreadsheet.

Table S2.10. List of expressed genes in the *W2* region according to the new reference genome (Daccord et al. 2017)

Table is provided as a supplementary spreadsheet.

Table S2.11. Differentially expressed genes (DEG) and genes of interest in the *W* and *W2* regions according to both versions of the apple reference genome.

In V2- the new reference genome (Daccord et al. 2017)										In V1-the first version of apple reference genome (Velasco et al. 2010 and Bai et al. 2014)	
Gene ID	Std pool (RPKM)	Weep pool (RPKM)	Fold Change (original values)	Baggerley's test: P-value	Baggerley's test: FDR p-value correction	Annotations	Chr region	Chr region start	Chr region end	Gene IDs[1]	Chromosomal locations and home contig IDs
MD13G1118800	160.3	76.8	-2.1	1.43E-06	1.17E-04	Rubber elongation factor protein (REF)	W-Chr13	8,684,773	8,686,689	M160372	chr13_8226278_8227987+_MDC000402.360
MD13G1119900	27.5	2.9	-9.6	1.17E-05	8.15E-04	Protein of unknown function (DUF579)	W-Chr13	8,781,443	8,784,582	M928608	chr13_8324364_8325212-_MDC002436.463
MD13G1120100	146.4	43.2	-3.4	1.72E-12	3.31E-10	Protein of unknown function (DUF579)	W-Chr13	8,807,010	8,807,859	M534197	chr13_8340268_8341116+_MDC000565.204
MD13G1119100	156.3	90.6	-1.7	4.10E-04	0.019	photosystem I light harvesting complex gene 2	W-Chr13	8,734,031	8,736,151	M309014	chr13_8265920_8269493-_MDC021754.182
MD13G1121000	191.2	98.3	-1.9	1.75E-06	1.41E-04	nuclear transport factor 2A	W-Chr13	8,890,264	8,892,830	M155347	chr13_8381572_8389769-_MDC000565.206
MD13G1123300	1.4	19.6	13.8	6.13E-05	3.59E-03	BCL-2-associated athanogene 5	W-Chr13	9,142,915	9,143,929	M153978	chr13_8675928_8677484+_MDC018786.82
MD13G1127100	12.2	101.1	8.3	1.61E-17	4.91E-15	jasmonate-zim-domain protein 10	W-Chr13	9,515,819	9,518,542	G105518	chr13_9555705..9579483_MDC015872.319
MD13G1127300	131.8	46.0	-2.9	2.31E-09	2.99E-07	acyl-CoA oxidase 1	W-Chr13	9,572,285	9,574,996	M670256	chr13_9627490_9629693+_MDC001448.405
MD13G1122400	32.0	23.9	-1.3	0.441	1	Protein of unknown function, LAZY1-like	W-Chr13	9,029,505	9,032,392	M254069	chr13_8556287_8569904-_MDC027667.11
MD13G1122900	8.6	4.4	-2.0	0.298	1	PCF (TCP)-domain family protein 20	W-Chr13	9,108,308	9,109,998	M374900	chr13_8629824_8630576-_MDC000204.221
MD10G1192900	146.8	208.3	1.4	5.10E-05	3.03E-03	indole-3-acetic acid 7	W2-Chr10	28,909,336	28,912,808	G104254, G104253	chr10_19596767..19598151_MDC011683.379
MD10G1199900	49.5	15.5	-3.2	7.65E-05	4.33E-03	glycine-rich protein	W2-Chr10	29,806,896	29,808,178	M142356	chr10_20510647_20511678+_MDC017711.154
MD10G1202900	37.1	9.8	-3.8	1.62E-04	8.49E-03	SBP (S-ribonuclease binding protein) family protein	W2-Chr10	30,057,823	30,059,373	M819881	chr10_20823655_20824964+_MDC009941.125
MD10G1196900	23.8	50.4	2.1	6.64E-04	0.029	glutathione S-transferase TAU 19	W2-Chr10	29,492,440	29,499,355	M178233	chr10_20273431_20274212+_MDC006660.401
MD10G1203300	137.1	184.4	1.3	6.83E-04	0.03	Ribosomal protein S5/Elongation factor G/III/V family protein	W2-Chr10	30,097,036	30,101,104	G102599, G102600, M777793	chr10_20832703..20870819_MDC007289.107

MD10G1196600	0.0	13.2	∞	2.75E-04	0.014	glutathione S-transferase TAU 25	W2-Chr10	29,460,982	29,461,586	NA	
NA										G102554	chr10_19740828..19807025_MDC007223.300
MD10G1193000	21.2	28.7	1.4	0.175	1	AUX/IAA transcriptional regulator family protein	W2-Chr10	28,924,799	28,926,049	M176753	chr10_19773424_19774161-_MDC007223.300
MD10G1202100	6.7	5.1	-1.3	0.742	1	SAUR-like auxin-responsive protein family	W2-Chr10	29,998,839	29,999,244	M186167	chr10_20781004_20781408-_MDC009941.128
MD10G1204800	6.2	1.1	-5.7	0.068	1	SAUR-like auxin-responsive protein family	W2-Chr10	30,376,679	30,377,099	M138076	chr10_21148059_21149432-_MDC010308.342
MD14G1073800							Chr14			G103289	chr10_19592795..19596965_MDC009423.544

[1] The prefix 'MDP0000' (e.g. MDP0000123456) in the original gene IDs is abbreviated to 'M' (e.g. M123456). Gene IDs in 'G######s' as in 'G101234' are for the novel transcripts reported in Bai et al (2014).

DEG by both V1 and V2 apple reference genomes.

DEG by V2.

DEG by V2, but outside of the W region in V1.

non-DEG, but a potential candidate gene by function

DEG by V1.

DEG by V1. But it is on chromosome 14 in V2.

Supplementary References

Bai Y, Dougherty L, Xu K. (2014). Towards an improved apple reference transcriptome using RNA-seq. *Mol Genet Genomics, 289*, 427-438.

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

CHAPTER 3

MdLazy1: a strong candidate gene for weeping growth habit in *Malus*

Abstract

In *Malus*, weeping growth habit is characterized by downward growing branches. Here, we report the identification of *MdLazy1*, a strong candidate gene under the *Weeping* (*W)* locus that largely controls the weeping growth trait in *Malus*. *MdLazy1* is an ortholog of *Lazy1-like* genes involved in branch angles and gravitropism in Arabidopsis, rice and maize. The weeping allele of *MdLazy1* contains a L195P substitution, which falls within a predicted transmembrane domain. Subcellular localization of the standard and weeping alleles revealed that they both localize at the plasma membrane and nucleus. Transgenic trees under expressing the standard *MdLazy1* allele had wider leaf and branch angles than those over expressing the allele. However, tree over expressing the weeping allele and *MdLazy1* RNAi lines had similar phenotypes as those under expressing the standard allele. These data suggest that the L195P mutation is likely causal for the weeping phenotype. Yeast two hybrid screening identified eleven interactors of *MdLAZY1* including brevis radix and regulator of chromosome condensation 1 genes.

Introduction

Branch angles are defined by the gravitropic set-point angle (GSA), which measures the angle of a given branch from gravity (Digby & Firn, 1995). In woody trees, shoots exhibit negative gravitropism growing upward against gravity, while roots exhibit positive gravitropism growing downward with gravity. Gravitropism can be broken down into three steps: perception, biochemical signaling and differential growth (Vandenbrink & Kiss, 2019). It is generally accepted that in gravity sensing cells called statoliths, starch accumulating amyloplast settles in response to gravity (Morita, 2010). This sedimentation causes a signaling cascade, ultimately leading to differential growth response (Chen, Rosen, & Masson, 1999). Auxin and strigolactone are key hormones in gravitropism (Moore, 2002; Sang et al., 2014). Uneven auxin distribution leads to shoot or root tip bending (Hart, 1990). While the exact mechanisms involved in gravitropic signaling are unknown, relevant genes are being uncovered (Friml, Wiśniewska, Benková, Mendgen, & Palme, 2002; Toyota & Gilroy, 2013; N. Zhang et al., 2018). These genes include a heat stress transcription factor (*HSFA2D*), a MADS-box transcription factor (*MADS57*) and an early phytochrome responsive transcription factor (*EPR1*) that work upstream of auxin response (N. Zhang et al., 2018).

In *Malus* (apple), the weeping growth habit is defined by downward growing branches and is commonly found in crabapple. While weeping crabapples are desirable for ornamental purposes, it is important to unearth the underlying mechanisms behind the growth habit. In a commercial orchard setting, apple trees are grown on trellis systems where the branches are manipulated to allow optimal fruit production. The espalier and spindle training systems bend branches outward to a horizontal position (Mika, 1992). Knowing the genetic mechanisms of branch angle can allow future manipulation of tree branches genetically instead of manually. In

apple, the weeping growth habit is a dominant trait, mostly controlled by the *W* (weeping) locus (Just, 2001; Sampson & Cameron, 1965). *W* was recently mapped to a 1MB region on chromosome 13 (Dougherty, Singh, Brown, Dardick, & Xu, 2018). In this study, eleven potential candidate genes of *W* were speculated based on differential expression between standard and weeping trees. They included glucuronoxylan 4-O-methyltransferase (GXMT)-like genes and a rubber elongation factor (Dougherty et al., 2018). Within the *W* locus, a *Lazy1-like* gene (*MdLazy1*) was also present although not differentially expressed. However *Lazy1*, a member of the IGT family is a well-characterized gene involved in shoot gravitropism (Dardick et al., 2013; Dong et al., 2013; P. Li et al., 2007; Taniguchi et al., 2017; Yoshihara, Spalding, & Iino, 2013).

Lazy1 has been primarily studied in *Arabidopsis thaliana, zea mays* and *Oryza sativa* (Dardick et al., 2013; Dong et al., 2013; P. Li et al., 2007; Z. Li et al., 2019; Sasaki & Yamamoto, 2015; Taniguchi et al., 2017; Wang et al., 2018; Yoshihara et al., 2013; J. Zhang et al., 2014). When *Lazy1* is functional, shoots have narrow angles and upright growth (Dardick et al., 2013). Mutant plants with reduced or no *Lazy1* expression have wide angle branches, downward growing branches and an overall weeping appearance (Hill & Hollender, 2019). In *Zea mays, Lazy1* mutants have a pronounced prostrate growth (Dong et al., 2013). In Arabidopsis, *AtLazy1* is involved in the signaling transduction pathways downstream amyloplast sedimentation (Taniguchi et al., 2017). In *O. sativa,* lateral auxin movement was inhibited in *Lazy1* mutants (Godbole, Takahashi, & Hertel, 1999; Yoshihara & Iino, 2007). A working model has been proposed where *Lazy1* function is before auxin redistribution and growth response but after amyloplast sedimentation (Hill & Hollender, 2019).

In this study we propose *MdLazy1* is a strong candidate gene under locus *W*. Characterization of *MdLazy1* suggests it is orthologous to the *Lazy1* genes described above. The

weeping allele of *MdLazy1* has a single nucleotide polymorphism $CT_{584}T>CC_{584}T$ in the coding sequence that results in a non-synonymous substitution L195P. Transgenic apple trees overexpressing the weeping allele has downward growing leaves and those overexpressing the standard allele had straight leaves and branches with narrow angles, suggesting that the L195P substitution is likely causal for the weeping growth habit. Yeast two hybrid screening identified 11 unique interactor of *MdLazy1* including regulator of chromosome condensation 1 (*RCC1*) and Brevis radix family proteins.

Materials and methods

Plant materials, phenotyping and fine mapping of *W* locus

A 'Cheal's Weeping' open pollinated (CW OP) population of 175 seedling trees segregating for weeping growth habit was used for fine genetic mapping of the *W* locus on chromosome 13. Trees were phenotyped as weeping (W), weeping-like (WL), standard (S) or standard-like (SL). Weeping trees had all branches growing downwards, while weeping-like trees had a majority of branches growing downward. Standard trees were classified by all upright branches and standard like trees had a majority of upright branches and overall upright structure. **(Figure S3.1).**

DNA was extracted from leaf tissue using a CTAB method previously described (Doyle, 1987). High resolution melting (HRM) markers targeting single nucleotide polymorphisms (SNPs) were developed within the previously described *W* region (Dougherty et al., 2018) **(Table S3.1).**

MdLazy1 promoter (C13SR_8547F1/MdLAZYPrF2) and *MdLazy1* SNP (lzySNP584HRMF/R) markers were developed and screened with CW OP and 35 F_1 progeny

from a ‘Cheal’s Weeping’ (weeping) × ‘Evereste’ (standard) cross and 124 F_1 progeny from a ‘NY-051’ (standard) × ‘Louisa’(weeping) cross previously used to identify the *W* locus (Dougherty et al., 2018). Additionally a segregating ‘Red Jade’ (weeping) OP population (RJ OP) of 226 progeny and six diverse weeping crabapple accessions that included ‘Molten Lava’, ‘Ludwick’, ‘A-23’, ‘Beverly NSY’, ‘Sinai Fire’ and ‘Ann E. Manback weeper’ from the F.R. Newman Arboretum, Cornell University, Ithaca, NY and two from the US National Malus germplasm repository, Geneva, New York, USA that included ‘Cascade’ and ‘Oekonomierat Echter-meyer’ were phenotyped and screened.

Determination of the genomic and cDNA sequences of *MdLazy1* alleles

The genomic DNA sequence of *MdLazy1* was amplified from ‘Cheal’s Weeping’, ‘Red Jade’, ‘Golden Delicious’, ‘NY-051’ and ‘Lousia’ with primers Lazy1F/R_S and cloned into pJET1.2/blunt vector (ThermoFisher Scientific, Waltham, MA). Sanger sequencing was conducted at Cornell University Biotechnology Resource Center (Ithaca, NY). Actively growing apical meristem tissues were sampled from ‘Cheal’s Weeping’ and were flash frozen in liquid nitrogen and stored under -80C until use. RNA was isolated from one gram of ground tissue as previously described (Meisel et al., 2005) with slight modifications. Briefly, ground tissues were mixed with CTAB buffer and incubated for 30 minutes at 65°C, then centrifuged. Supernatant was mixed with equal amounts chloroform and centrifuged again. RNA was precipitated from the resulting supernatant in lithium chloride at -20°C for 2 hours and resuspended in 87.5ul of water. The RNA samples were treated with DNase I (amplification grade, Invitrogen, Carlsbad, CA) and cleaned with RNeasy MinElute Clean up Kit (Qiagen, Hilden, Germany). RNA concentrations were determined using NanoDrop 1000 (Thermo Fisher Scientific) and RNA

integrity was assayed on 1% agarose gel. One microgram of total RNA was used in reverse transcription reactions using the Superscript III RT (Invitrogen, Carlsbad, CA, USA) to obtain the first strain cDNA.

The CDS of *MdLazy1* standard allele *(MdLazy1*-S*)* and *MdLazy1* weeping allele *(MdLazy1*-W*)* with and without the stop codons were PCR-amplified from 'Cheals Weeping' cDNA using primers Lazy1F/R_S and Lazy1F/R_NS, respectively **(Table S3.1)**. The PCR products were cloned into Gateway entry vector pCR8/GW/TOPO (Invitrogen) and were sequence-confirmed.

MdLazy1 promoter construct and Gus assay

A 1.69-kb fragment upstream the *MdLazy1-W* ATG start codon from 'Cheal's weeping' was PCR amplified using LA taq (Takara Bio Inc, Kusatsa, Shiga Prefecture, Japan) according to manufacturer's protocol and then cloned into Gateway entry vector pCR8/TOPO/GW (Invitrogen, Carlsbad, CA, USA). A sequence confirmed clone was transformed into the ß-glucuronidase (Gus) reporter vector pMDC164c (Curtis & Grossniklaus, 2003) via LR reaction (Invitrogen). The completed construct called MdLazy1WPro was transfected into Agrobacterium strain EHA105. *Arabadopsis thaliana* (Columbia) plants were transformed by the floral dip method (X. Zhang, Henriques, Lin, Niu, & Chua, 2006). Seeds were collected from dipped plants, sterilized and plated on MS selection plates containing hygromycin B. Seeds were placed in the dark at 4°C for three days then exposed to light for six hours to promote germination and then returned to the dark at room temperature for five days. The seedlings that grew vertically after five days in the dark were considered transgenic and transferred to soil. At their different development stages, the plants were stained with a β-Glucuronidase Reporter Gene Staining Kit

(Sigma-Aldrich, St. Louis, MO, USA) according to manufacturer's protocol. Images were acquired using an Infinity2 camera (Lumenera, Ottawa, ON, Canada) on a Stemi 2000-C stereoscope (Zeiss, Oberkochen, Germany).

Subcellular localization of *MdLazy1*-S and *MdLazy1*-W

For subcellular localization, *MdLazy1*-S and *MdLAZY1*-W clones (without stop codon) in pCR8/TOPO/GW vectors were used to transfer the CDS into Gateway compatible green fluorescent protein (GFP) fusion vector pEarleyGate 103 (Earley et al., 2006) by LR reaction. The resultant constructs, named *MdLazy1*-S:PEG103 and *MdLazy1*-W:PEG103, were transformed into Agrobacterium (Strain GV3101::pMP90) by electroporation. Positive clones were grown overnight to OD_{600}=1. The agrobacterium was resuspended in 10mM MES, 10mM MgCl2, 200uM acetosyringone and injected into 4-week old *Nicotiana tabacum* leaves by needleless syringe and grown in normal conditions for 3-7days. Leaf pieces were observed under an Olympus FV3000 (IX83) confocal microscope (Olympus Corporation, Shunjuku, Tokyo, Japan) with 488 nm and 633 nm lasers and 40x objective. *MdLazy1*-S:PEG103 was co-injected with plasma membrane marker CD3-1007 (mCherry) (Nelson, Cai, & Nebenfuhr, 2007) as described above and observed with 488nm, 587nm and 633nm lasers.

***Lazy1-S* and *Lazy1-W* overexpression constructs**

The *MdLazy1*-S and *MdLAZY1*-W with stop codon clones in pCR8/TOPO/GW vectors were used to transfer the CDS into the plant overexpression vector pGWB412 by LR reaction (Invitrogen) and were sequence-confirmed, resulting in constructs *MdLazy1*-S:pGWB412 and *MdLazy1*-W:pGWB412, respectively.

MdLazy1 RNAi plasmid construction

A *MdLazy1* RNAi construct was created following previously described guidelines (C. Helliwell & Waterhouse, 2003). Briefly, a 459bp gDNA fragment of *MdLazy1* with flanking attB sites was PCR amplified with primers attB1-Lazy1-F/attB2-Lazy1-R (Table S1), cloned into in pCR8/TOPO/GW vector (Invitrogen) and sequenced for confirmation. Clones of correct sequence were then transferred into pHELLSGATE2 RNAi vector (C. A. Helliwell, Wesley, Wielopolska, & Waterhouse, 2002) by BP reaction (Invitrogen), creating the *MdLazy1*-RNAi plasmid. *MdLazy1*-RNAi was digested with Xbal (New England Biolabs, Ipswich, MA and Xhol (New England Biolabs) for 2 hours at 37C and run on 1.5% agarose gel to confirm insert size. Sequencing of *MdLazy1*-RNAi was completed with primers P27-5 and P27-3 (**Table S3.1**).

MdLazy1-S*:*pGWB412*, MdLazy1*-W*:*pGWB412 and *MdLazy1*-RNAi apple transformation

For apple leave transformation, *A. tumefaciens* strain EHA105:pCH32 containing the constructs *MdLazy1*-S:pGWB412, *MdLazy1*-W:pGWB412 and *MdLazy1*-RNAi were prepared as previously reported ((Borejsza-Wysocka, Norelli, Aldwinckle, & Ko, 1999); Norelli et al. 1999). Leaves used in transformation were harvested from 'Royal Gala' strain (*MdLazy1*-S*:*pGWB412) and 'Royal Gala' (*MdLazy1*-W:pGWB412 and *MdLazy1*-RNAi) leaf expansion cultures that were actively growing. The leaf selection for harvest was the youngest unfolded, but still expanding, leaf on the shoot apex. The entire leaf surface was wounded using non-traumatic forceps. Immediately after wounding, leaves were placed, abaxial side-down, in a petri dish containing the *A.tumefaciens* inoculum. Leaves were remained in the inoculum for 5 minutes. Excess inoculum was removed from leaves by blotting on sterile filter paper. Leaves were then transferred abaxial side-down on cocultivation medium with 100μM acetosyringone and 1 mM

betaine phosphate, and incubated for 3 days in the dark at 25°C.

After 3 days leaves were removed from cocultivation medium and washed with sterile half-strength Murashige and Skoog salt mixture plus 500µg cefotaxime for 10 minutes. After leaves were blotted dry on sterile filter paper and transferred to moist chamber. The leaf tip and petiole were removed from the leaf using a sterile scalpel and remaining leaf blade was cut into 3 strips (ca. 2-3 mm wide). Leaf pieces were transferred abaxial side up to regeneration medium containing 100µg kanamycin and 350µg cefotaxime/ml. Plates were incubated in dark at 25°C for three weeks. After three weeks of incubation plates were transferred to 16 h lights and 25°C. Single kanamycin-resistant regenerants were transferred to proliferation medium containing 100µg paramomycin and 250µg cefotaxime/ml. Established lines were propagated and rooted.

Analysis of Transgenic *MdLazy1*-S trees and *MdLazy1*-W trees

DNA was extracted from leaves of all transgenic (*Lazy1*-S:pGWB412/*Lazy1*-W:pGWB412) and wild type ('Royal Gala' strain or 'Royal Gala') trees as described above. Primers 35SF1/Lazy-1R (Table S1) were used to confirm the transgene presence or absence. For *MdLazy1*-S:pGWB412 trees (18 months old) actively growing apical meristem, branch shoot tips and leaf petioles were flash frozen in liquid nitrogen for RNA isolation and cDNA synthesis as described above. cDNA was used as templates for Quantitative PCR (qPCR) analysis of *MdLazy1* expression. qPCR reactions were performed with three replicates using iTaq Univeral SYBR Green Supermix (BioRad, Hercules, CA) on a CFX96 Touch Real-Time PCR Detection System (BioRad) according to manufacturer's protocol. An apple actin encoding gene (EB136338) was used as a reference gene **(Table S3.1).** The expression levels of *MdLazy1* genes were quantified based on the normalized expression (ΔΔCq) of the reference gene actin using the

Bio-Rad CFX Maestro software. For the *MdLazy1*-W:pGWB412 transgenic trees (8 months old), apical shoot tips were collected and analyzed similarly for *MdLazy1* expression.

Quantitative measurements for branch angle and branch tip orientation were recorded for *MdLazy1*-S:pGWB412 trees. Overall tree appearance and leaf orientations were noted. For *MdLazy1*-W:pGWB412 trees (six months old), leaf petiole angles were measured and leaf orientation was noted. Trees did not have branches. All measurements are assuming that 0° is straight up, 90° is horizontal and 180° is straight down, which is opposite from previous papers (Digby & Firn, 1995; Roychoudhry, Del Bianco, Kieffer, & Kepinski, 2013).

RNA sequencing

Apical shoot tips from eight 'RJ' OP samples, 'Cheal's Weeping' and 'Evereste' were collected, flash frozen and total RNA was extracted, purified and quantified as described above. mRNA was isolated with NEBNext Poly(A) mRNA Magnetic Isolation Module and used to create RNA-seq libraries with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) according to manufactures protocol with slight modifications. 10mg of total was used as for mRNA isolation input and only half the isolated mRNA was used for library preparation. Libraries were multiplexed in equal amount for single end 101-base sequencing by HiSeq 2500 (Illumina, San Diego, CA) at the Cornell University Biotechnology Resource Center (Ithaca, NY). RNA-seq data analysis was conducted using CLC Genomic workbench 12.0 (Qiagen).

Yeast two hybrid screening

Y2H screening was carried out using the ProQuest Two Hybrid system (Invitrogen)

according to manufactures protocol. Briefly, *MdLazy1*-S was used as bait. A custom tree architecture prey library made from 'Cheal's Weeping', 'Red Jade', 'McIntosh' and 'Wijcik McIntosh' apical shoot tips using CloneMiner II cDNA Library Construction Kit (Invitrogen) and used for screening. After initial screening, any positive interactors were sequenced and co-transformed with *MdLazy1* for confirmation in yeast.

Protein prediction software and protein alignment

The coding sequences (CDS) of *MdLazy1*-S and *MdLazy1*-W were amplified from 'Cheal's Weeping' cDNA and sequenced. The protein sequences were screened with LOCALIZER (Sperschneider et al., 2017) and TMpred (Hofmann & Stoffel, 1993) programs to identify nuclear localization signals or transmembrane domains respectively.

Protein sequences from *Lazy1* gene homologs in Arabidopsis (Q5XV40), Rice (Q2R435) and Maize (B4FG96) were downloaded from UniProt (UniProt Consortium, 2018) and aligned with the protein sequences of *MdLazy1*-S and *MdLazy1*-W sequences using ClustalΩ SnapGene® software (from GSL Biotech; available at snapgene.com).

Statistical analysis

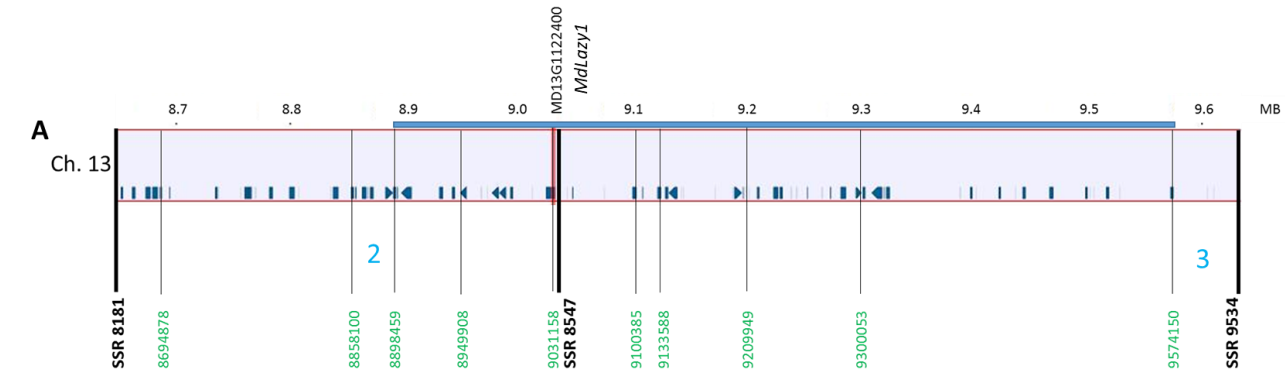
Various statistical analyses, including one-way analysis of variance (ANOVA), Tukey-Kramer HSD, correlation and regression analyses were conducted using software package JMP Pro 14 (SAS, Cary, NC, USA). Significance was determined by $p < 0.05$.

Results

Phenotyping and Fine Mapping

The ‘Cheal’s Weeping’ OP population segregated at a 1:1 ratio of weeping (W/WL) to standard (S/SL) ($\chi 2$=0.966, P=0.325). Additionally RJ OP ($\chi 2$=0.15, P=0.698) segregated at a 1:1 ratio **(Figure S3.2).**

To fine map *W*, new high resolution melting (HRM) markers were designed within the 982 kb region defined by markers SSR8181 and SSR9530 (Dougherty et al., 2018)(**Table S3.1).** The HRM markers were based on single nucleotide polymorphisms (SNPs) in genomic DNA where the ‘Cheal’s Weeping’ parent was homozygous for the polymorphism compared to the reference genome (Daccord et al., 2017). Three informative recombinants were found in the ‘Cheal’s weeping’ OP progeny and sequence confirmed for the SNPs, narrowing down the *W* locus to 675 kb that contained 60 predicted genes **(Figure 3.1).** Additionally, two progeny from the ‘NY-051’ × ‘Louisa’ population were recombinant.



B

Recombinants	8898459	9031158	Phenotype	9100385	9574150
166	G/G	T/T	S	T/T	G/G
170	G/G	T/T	S	T/T	G/A
144	G/A	T/T	S	T/T	G/G
169	G/A	T/C	WL	T/C	G/G
Cheal's Weeping	G/A	T/C	W	T/C	G/A

Figure 3.1. Fine mapping of *W* region and location of recombinants. **A.** Overview of *W* region on chromosome 13 with predicted genes and chromosomal locations (MB). Bold black markers indicate SSR markers previously defined the *W* region. Newly developed HRM markers are shown in green. The number of recombinant plants from CW OP and 'NY-051' × 'Lousia' progeny for each marker are shown in blue. The newly derfined *W* region spanning the blue bar is 0.675MB. **B.** Map of CW-OP HRM marker recombinants. Top row indicated the position of the SNP the HRM marker is targeting while the first column shows the plant ID. Nucleotides at SNP location are indicated. Plant 166 was used as a standard control and 'Cheal's Weeping' was used for the weeping control. Recombinants are based on SNP comparison with 'Cheal's Weeping'. Phenotypes are Standard (S), Weeping like (WL) and Weeping (W).

MdLazy1 alleles and *MdLazy1* specific markers

Sequencing of the *MdLazy1* alleles revealed a 62 bp deletion in the promoter region (692bp upstream the start codon) in 'Cheal's Weeping' (weeping phenotype), 'Red Jade' (weeping phenotype), and 'Lousia'(weeping phenotype) but not in 'NY-051'(standard phenotype) or 'Golden Delicious' (standard phenotype) **(Figure 3.2 A).** To see if the promoter deletion was linked to the weeping phenotype, SSR marker C13SR_8547F1/ MdLAZYPrF2 was designed flanking thedeletion region. The target size of the marker was 200bp based on the reference genome. When run, a weeping allele, with deletion, was indicated by a band size of

approximently 163bp while bands indicating standard alleles can ranged from 200 bp to 260bp in length **(Figure 3.2 B)**. There a multiple microsatailites present in the amplified reigon, leading to size differences observed. This marker was screened with all populations and the eight diverse weeping varieties, uncovering that the deletion is tightly linked to the weeping phenotype: 'Cheal's Weeping' OP ($R^2 = 0.649$, $p < 0.00001$), 'Red Jade' OP ($R^2 = 0.88$, $p < 0.00001$), 'Cheal's Weeping' x 'Evereste' ($R^2 = 0.708$, $p < 0.0001$), and 'NY-051' × 'Louisa' ($R^2 = 0.88$, $p<0.0001$). Additionally all six arboretum weeping varieties and two germplasm weeping varieties had the promoter deletion

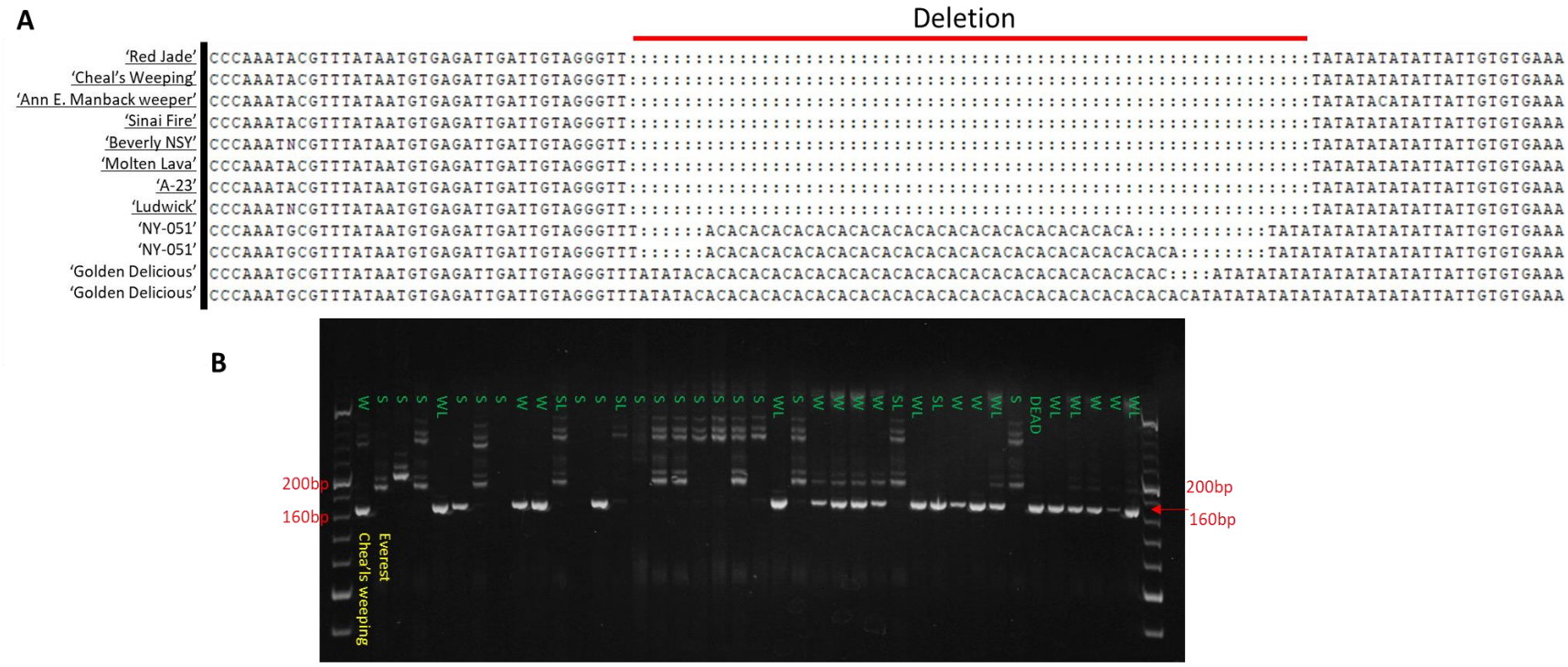


Figure 3.2. Identification of *MdLazy1* promoter deletion**. A.** Sequences of weeping and standard apple trees at the promoter deletion region 692bp upstream the start codon. Weeping varieties (underlined) have a 62bp deletion. Standard trees had no deletions or smaller deletions (4-20bp) **B.** SSR marker C13SR_8547F1/MdLAZYPrF2 gel profile. Deletion bands are indicated by the blue arrow above the 160bp ladder marker. 'Cheal's Weeping' and 'Evereste' are used as weeping and standard controls. Tree phenotypes are listed as 'W' (weeping), 'WL' (weeping like), 'S' (Standard) and 'SL' (Standard like) above each lane

Investigation of the *MdLazy1* CDS revealed no structural differences such as deletions/insertions between weeping and standard progeny that would introduce frameshift mutations, however there were two SNPs that led to non-synonymous changes. The first was at $C_{250}AT>T_{250}AT$ resulting in a histidine to tyrosine change at amino acid (AA) 84 (H84Y). 'Cheal's weeping' was heterozygous for C/T, while 'Red Jade' and 'Golden Delicious' were homozygous C/C, suggesting this SNP is not linked to weeping phenotype. The second non-synonymous change was $CT_{584}T>CC_{584}T$ resulting in leucine to proline change at AA 195(L195P). Amplification and sequencing of the CDS from 'Cheal's weeping' indicated there are two alleles present **(Figure S3.3**). One allele had a 'T' at base 584 while the other had a 'C'. Revisiting the gDNA sequence, the weeping varieties were heterozygous for T_{584}/C_{584} and the standard are homozygous for T_{584}/T_{584}. We refer to the $CT_{584}T$ allele as *MdLazy1*-S since it is the same as the 'Golden Delicious' (standard) reference sequence and the $CC_{584}T$ allele as *MdLazy1*-W because $CC_{584}T$ was found in weeping varieties only. Based on this information we designed a HRM marker (lzySNP584HRMF) to target the $T_{584}T>CC_{584}T$ SNP and determine if it was linked to the weeping phenotype. In the 'Cheal's Weeping' OP population, the SNP could explain most of the variations ($R^2 = 0.7088$, $p < 0.0001$). Further screening with 'Cheal's Weeping' x 'Evereste' ($R^2 = 0.561$, p= 4.5E-4), and 'NY-051' × 'Louisa' ($R^2 = 0.88$, $p <0.0001$) showed strong correlation to the weeping phenotype.

'Red Jade', 'Cheal's Weeping', and 'Lousia' were sequence confirmed to be heterozygous for $CT_{584}T/CC_{584}T$ as were the diverse weeping samples **(Figure S3.4).** Standard 'Golden Delicious', and 'NY-051' were sequence confirmed to be homozygous for $CT_{584}T$.

MdLazy1 promoter Gus assay

A 1.69Kb promoter region upstream of ATG codon of *MdLazy-W* was amplified from 'Cheal's weeping' and sequenced **(Figure S3.5)**. Arabidopsis were stained at different developmental stages and Gus reporter activity was observed. In young Arabidopsis, expression was seen in the rosette leaves and the tap root. In older Arabidopsis, expression was seen in flowers, fruit, roots, leaves, trichomes and stems **(Figure 3.3).** The stems had bands of expression, with areas of darker and lighter staining. Darker staining is also observed at leaf junctions with petioles.

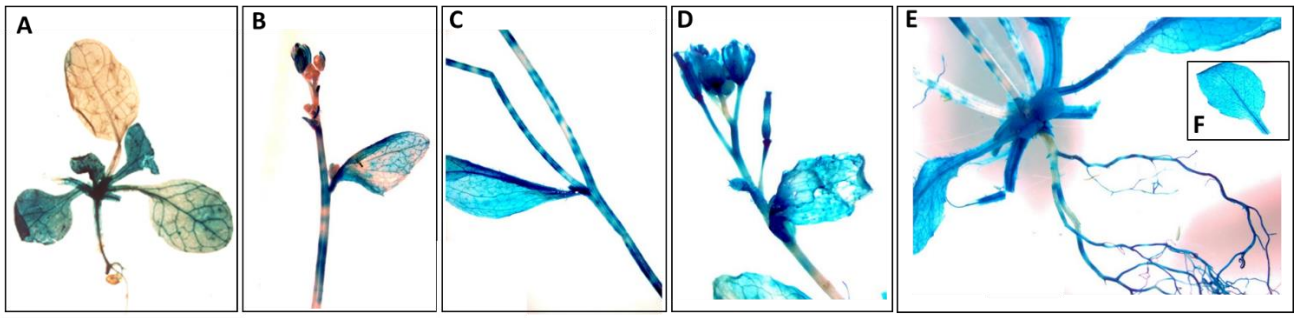


Figure 3.3. Expression of *MdLazy1-W* promoter assay. **A.** Young Arabidopsis plant, **B-F.** Older Arabidopsis plant **B.** flower buds, leaf junction **C.** Stem, internode and leaf junction, **D.** fruit and flowers, **E.** Rosette leaves, stems and roots, **F.** close up on rosette leaf.

MdLazy1 protein subcellular localization

MdLazy1-S and *MdLazy1*-W alleles were predicted to have nuclear localization signals using LOCALIZER. In TMpred transmembrane screening, the parameters were set for transmembrane helix length of 14-35AA. *MdLazy1*-S had one predicted transmembrane domain at amino acids 183-200 while no transmembrane domains were predicted for *MdLazy1*-W. Subcellular localization in tobacco, however, showed both alleles localized to cell nuclei and plasma membrane (**Figure 3.4, Figure S3.6**).

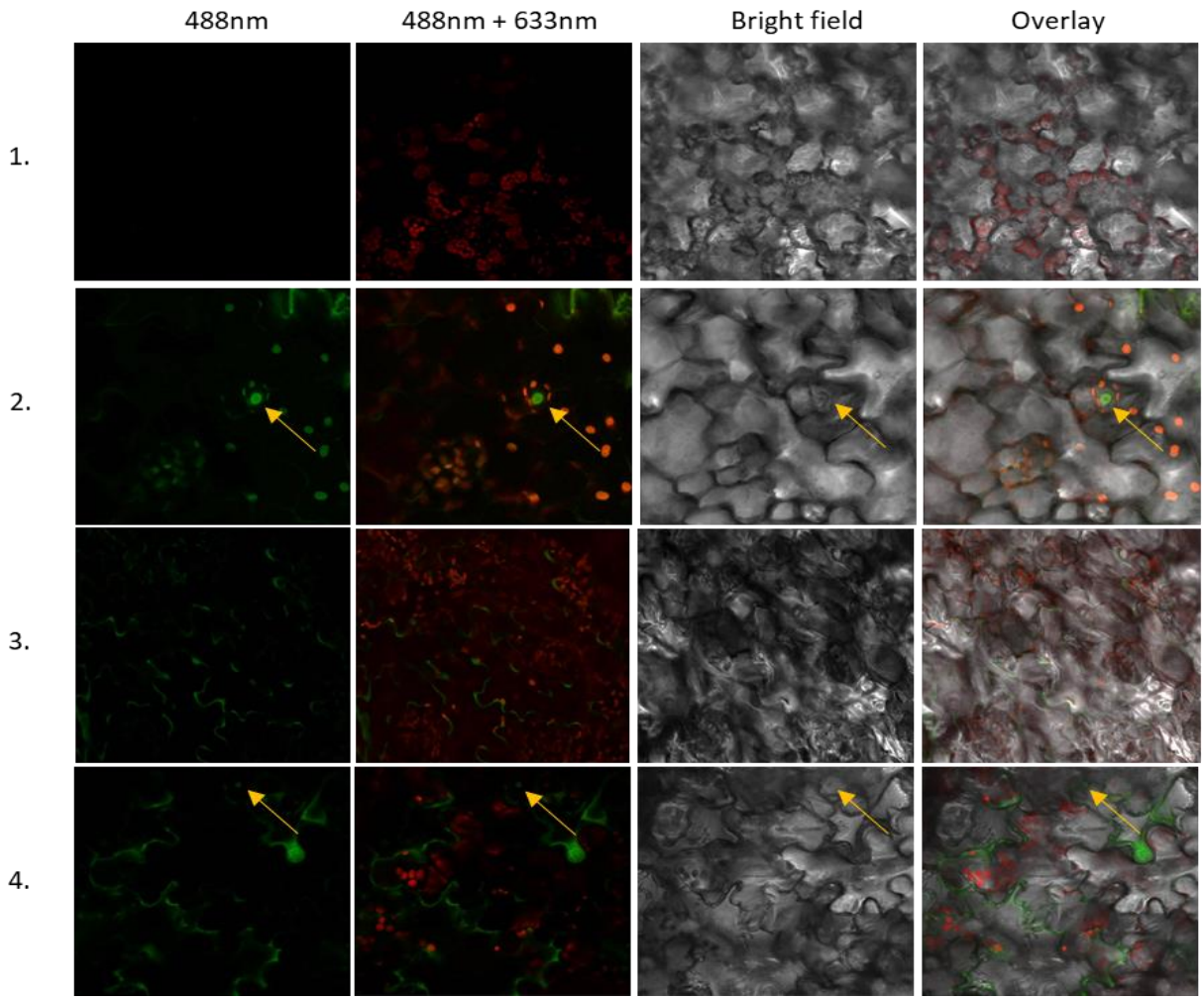


Figure 3.4. Subcellular localization of *MdLazy1* alleles in tobacco leaves**.** Row 1 shows negative control tobacco. Rows 2 and 3 show *MdLazy1*-S:GFP in the nucleus and plasma membrane respectively. Row 4 shows *MdLazy1*-W:GFP in the nucleus and plasma membrane. Columns indicate laser excitation (488nm= GFP, 633nm = chlorophyll). Orange arrows indicate nuclei.

MdLazy1-S, *MdLazy1*-W and *MdLazy1*-RNAi Transgenic apple trees

The *MdLazy1*-S:pGWB412 construct was transformed into leaves of GL3 (a 'Royal Gala' progeny of standard growth habit) by agrobacterium. Six independent lines were generated with multiple clones each line. The *MdLazy1*-W:pGWB412 construct was transformed into leaves of 'Royal Gala' generating seven independent lines (Lines 3, 4, 5A, 5B, 5D, 5E and 5F) with multiple replicates. Plants were grown in greenhouse under 16 hour light, 8 hour dark cycle

at 70°F.

MdLazy1-S:pGWB412 lines 1 and 7 displayed weeping like leaves from a young age, while line 6 and the GL3 control had upward leaves **(Figure S3.7)**. RNA isolation and subsequent qPCR for *MdLazy1* expression on leave petioles revealed lines 1 and 7 had reduced *MdLazy1* expression compared to the wild type control, while line 6 was over expressed **(Figure S3.7)**. At 18 months in age, a 40cm interval atthe top of each tree was defined and marked. Within that 40cm intervalbranch angle and tip orientation were measured for each primary branch **(Figure 3.5 A, B, C)**. The actively growing apical meristems and three branch tips were taken for qPCR analysis of *MdLazy1* expression **(Figure 3.5 D, E)**. The expression profiles were similar to the petiole samples, with Line 6 over expressing *MdLazy1* compared to the wild type control and line 1 under expressing *MdLazy1*. Line 7 apical shoot tips were slightly overexpressing *MdLazy1* when compared to wild type.

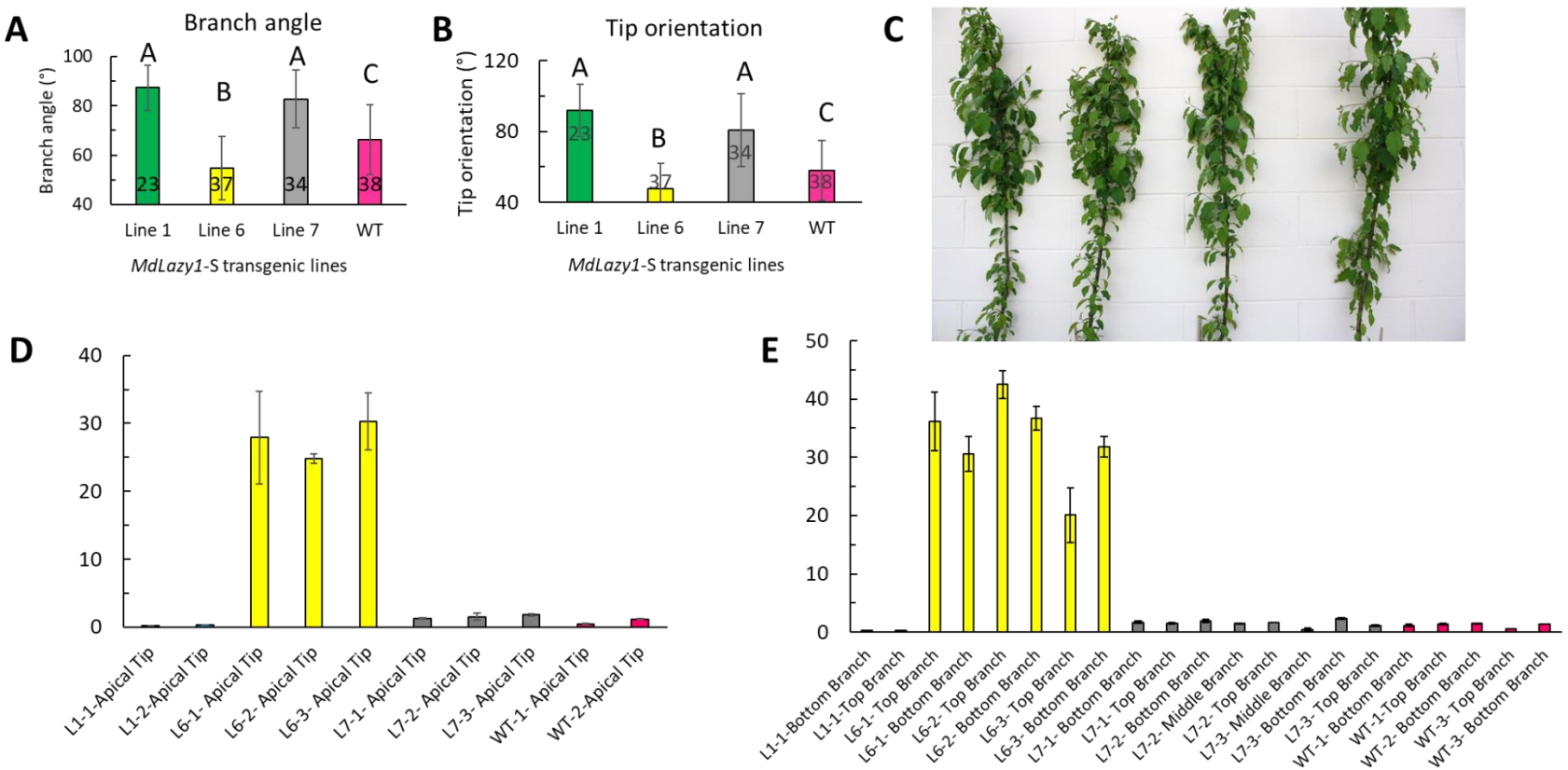


Figure 3.5. *Lazy1*-S:pGWB412 transgenic trees. **A.** Average branch angles for lines 1, 6, 7 and 'GL3' 'Royal Gala' strain (WT). **B.** Average tip orientation. Number of branches measured are shown in base of bars. Error bars are standard deviation. Significance is defined as $p < 0.05$. Different letters indicate significance between lines. **C.** Transgenic and WT trees with young branches, from left to right: line 7, line 6, WT and line 1. Relative *MdLazy1* expression in *MdLazy1*-S:pGWB412. **D.** Apical shoot tips, **E.** Branch tips. Individual lines are colored as follows: line 1 (blue), line 6 (yellow), line 7 (gray), WT (pink). Error bars are standard deviation.

MdLazy1-W:pGWB412 lines were five months old when their leaf angle measurements were taken. The plants were 130cm tall on average. As the plants grew, more extreme leaf angles are observed in the older leaves **(Figure 3.6 A).** To avoid a bias, a 30cm region 15cm from the top of the tree and a 30cm region 30cm from the soil were marked and all leaves within those regions were measured. For each tree 18+ leaves were measured, then the averages were calculated (**Figure 3.6 B**). Apical shoot tips were also taken from the trees for *MdLazy1* expression analysis. All the transgenic weeping lines had significantly higher expression of *MdLazy1* than the 'Royal Gala' control (**Figure 3.6 C).** Only one line of *MdLazy1*-RNAi trees was available at the time this chapter was written. The *MdLazy1*-RNAi line had downward leaves like those seen in *MdLazy1*-W over expression and *MdLazy1*-S under expression lines (**Figure S3.8**). Three replicates of *MdLazy1*-W:pGWB412 lines had their apical shoot tips removed to promote branching. Two months later, branch angles and shoot tip orientations were measured. The lines over expressing *MdLazy1*-W:pGWB412 had wider branch angles and tip orientations compared to 'Royal Gala' (**Figure S3.9**).

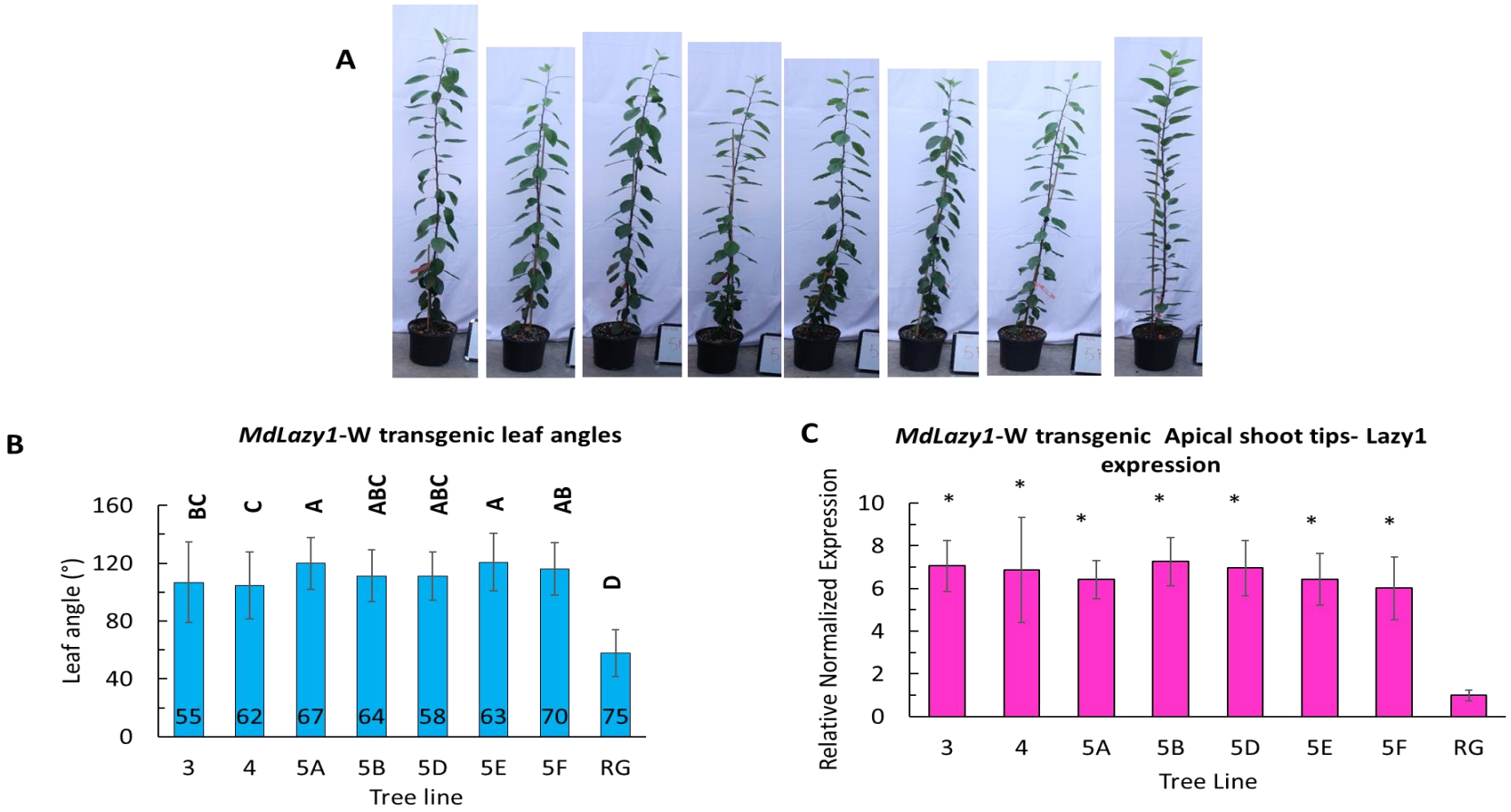


Figure 3.6. *MdLazy1*-W:pGWB412 transgenic 'Royal Gala' trees: **A.** Images of transgenic lines and 'Royal Gala' (RG) wild type control. From left to right, Line 3, 4, 5A, B, D, E, F and RG. **B.** Average leaf angles per line. Measurements were taken for all leaves within a 30cm region starting 15cm below the tip and a 30cm region starting 30 cm above the soil line. Three trees were measured for each line. The total number of leaves measured is shown on the graph. Significance is defined as $p < 0.05$. Different letters indicate significance between lines. **C.** Expression analysis of *MdLazy1* from apical shoot tips. All transgenic lines had similar leaf angles and had higher expression *MdLazy1* compared to the RG (royal Gala) control $p<0.05$.

Protein alignment

All proteins varied in size: apple 384AA, rice 416AA, maize 413AA and Arabidopsis 354AA. There were five conserved regions **(Figure 3.7)**.

Figure 3.7. Protein alignment of known *Lazy1* sequences and *MdLazy1* sequences using clustralΩ**.** Amino acids highlighted in yellow are highly conserved (AA is in 75% of proteins). Five main regions including the N terminus, IGT motif and CCL C-terminus are highly conserved. Protein names are on the right and amino acid positions are on the left. All four proteins differ in overall length. Above the consensus sequence: '*' single and fully conserved residue, ':' strongly similar conservation (>0.5 on PAM250 matrix), ' .' weakly similar conservation (< 0.5 on PAM250 matrix). The black arrow indicates the L→P mutation in *MdLazy1*-W which falls in a highly conserved region.

Identification of MdLAZY1-S protein-protein interactors by Y2H

Yeast two hybrid screening identified 43 positive interactors with *MdLazy1*-S allele. Sequencing of the interactors revealed 11 unique interactors (**Table 3.1)**. The 11 unique interactors consisted of five proteins in the regulator of chromosome condensation family (RCC1), three Brevix Radix like proteins and one NAD kinase, adenosine kinase 2 and DNase-I super family respectively.

Table 3.1 *MdLazy-1S* yeast two hybrid protein interactors: List of interactors confirmed in Y2H screening. Gene IDs correspond to apple genome reference genome GDDH3

Name	Gene ID	Annotations
Brevis Radix like 4	MD02G1085600	Myosin family protein with Dil domain
Brevis radix	MD06G1014400	Disease resistance/zinc finger/chomosome condensiton like region (DZC) domain
Brevis Radix like 2	MD09G1186000	Disease resistance/zinc finger/chomosome condensiton like region (DZC) domain
RCC1	MD03G1167700	Regulator of chromosome condenstation (RCC1) family with FYVE zinc finger domain
RCC1	MD07G1038100	Regulator of chromosome condenstation (RCC1) family with FYVE zinc finger domain
RCC1	MD07G1038400	Regulator of chromosome condenstation (RCC1) family with FYVE zinc finger domain
RCC1	MD10G1021300	Regulator of chromosome condenstation (RCC1) family with FYVE zinc finger domain
RCC1	MD11G1186800	Regulator of chromosome condenstation (RCC1) family with FYVE zinc finger domain
NAD kinase 1	MD03G1172500	nicotinamide adenine dinucleotide 1 kinase
Adenosine kinase 2	MD12G1116000	
Dnase-I like super family protein	MD15G1144700	

Discussion

MdLazy1 is a strong candidate gene under *W*

The populations used in this study segregated phenotypically at a 1:1 ratio (weeping: standard) to standard phenotype as expected for a single gene dominant trait. Fine mapping of the *W* locus in the 'Cheal's Weeping' OP population narrowed the region to 675KB region **(Figure 3.1).** In the previously defined *W* locus (982kb), eight differently expressed genes

between standard and weeping RNA pool were identified (Dougherty, Singh, Brown, Dardick, & Xu, 2018). In the newly defined 675KB region, only three of those genes: MD13G1123300 (BCL-2 associated athanogene 5), MD13G1127100 (*jasmonate-zim-domain protein 10*) and MD13G1127300 (*acyl-CoA-oxidase 1*) remained. BCL-2 associated anthanogene 5 (*BAG5*) has not extensively been studied. But in *Abrabidopsis*, *AtBAG5* is predicted to localize to the mitochondria, is strongly induced by abscisic acid and may coordinate stress induced hormones (Doukhanina et al., 2006). *Jasmonate-zim domain protein 10* is involved in plant defense and *acyl-CoA-oxidase 1* is involved in fatty acid beta oxidation (Arent, Pye, & Henriksen, 2008; Chung et al., 2008). Based on these studies, none of the differently expressed genes were good candidate genes for weeping growth habit in apple. Further scrutiny of the *W* locus revealed MD13G1122400, a Lazy1-like gene, which has been extensively studied in Arabidopsis, rice and corn and is involved in shoot gravitropism (Dong et al., 2013; P. Li et al., 2007; Yoshihara, Spalding, & Iino, 2013) therefore making it a strong candidate for *W*.

Subcellular localization of *MdLazy1*-S:GFP and *MdLazy1*-W:GFP were observed at the plasma membrane and nucleus in tobacco leaf tissue (**Figure 3.4**, **Figure S3.6**). These results are in agreement with *Lazy1* localization studies in maize, rice and *Arabidopsis* (Dong et al., 2013; P. Li et al., 2007; Yoshihara et al., 2013; Zhang et al., 2014). The localization to the nucleus was also predicted by LOCALIZER and a NLS was identified. In *A. thaliana*, there are six *Lazy1* genes, but only *AtLazy1* localizes to both the nucleus and plasma membrane (Taniguchi et al., 2017). Rice, maize, *Arabidopsis* and *MdLazy1*-S were all predicted to have a single pass membrane. Rice and maize Lazy1 have the same predicted transmembrane region, while Arabidopsis and apple differed. *MdLazy1*-W did not have a predicted transmembrane domain. Protein alignment of the *Lazy1* sequences showed five conserved regions. Notably the N-

terminal domain, the IGT motif (GϕL(A/T)IGT) (Dardick et al., 2013) and the CCL (conserved C terminus in *LAZY1*) formally called an EAR motif (Taniguchi et al., 2017). The IGT motif is a signature of IGT family genes which include *TAC* and *Lazy* (Dardick et al., 2013).

Analysis of the *MdLazy1* promoters

Three main structures of the *MdLazy1* promoter sequences were found (**Figure 3.2 A**). The first structure contained a 62bp deletion that is tightly linked to the weeping phenotype in populations studied and the diverse weeping cultivars, but not exclusively found in weeping trees. The second structurefound in 'Golden Delicious' had nodeletion. The third structurehad minor deletions of varying length (4-20bp). Despite these promoter structural differences there weres no difference in *MdLazy1* expression between standard and weeping trees in RNA seq analysis of 'Red Jade' OP trees (Fold change 1.29, p-value 0.25), nor was expression differences observed in the previous study with 'Cheal's Weeping' × 'Evereste' standard/weeping progeny (Fold change -1.3, p-value 0.44) (Dougherty et al., 2018).

The GUS promoter assay revealed the *MdLazy1*-W promoter was expressed not only in shoots, stems, inflorescences, and leaves, but also in the roots. In Arabidopsis, *AtLazy1* had very weak promoter expression in the roots, while *AtLazy2*, *AtLazy3*, *AtLazy4* and *AtLazy5* had strong expression (Taniguchi et al., 2017; Yoshihara & Spalding, 2017). *AtLazy1* was also expressed in the stems and vascular tissue, while *AtLazy4* was in the leaves and *AtLazy6* was in the petioles. *A. thaliana* has six identified Lazy1 genes, while apple only has two (Wang et al., 2018; Yoshihara & Spalding, 2017). In apple, *MdLazy1* may have multiple functions involving both shoot and root gravotropism as *MdLazy1* splice variants were also identified in this study (data not shown).

MdLazy1 $CT_{584}T>CC_{584}T$ mutation is likely causal for weeping phenotype in *Malus*

Previous studies have shown that *Lazy1* loss of function mutations led to larger branch angles and prostrate growth consistent with weeping phenotype (Dong et al., 2013; P. Li et al., 2007; Yoshihara et al., 2013). *Lazy1* RNAi silencing in plum also led to wider branch angles (Hill & Hollender, 2019). In contrast, over expression of *Lazy1* in poplar caused more narrow branch angles(Xu et al., 2017). Characterization of the maize mutant lazy plant1 (*la1*) revealed three different alleles of *Lazy1* (*ZmLA1*) called *la1-1*, *la1-2* and *la1-3*. In each allele an insertion or SNP led to a frameshift (*la1-1*), premature stop codon (*la1-2*) or splice site disruption (*la1-3*), but they developed the same prostrate growth habit (Dong et al., 2013). Interestingly in another maize mapping experiment, the same *ZmLA1* gene with different mutations (called *ZmCLA4*) controlled leaf angle (Zhang et al., 2014). In *ZmCLA4* RNAi knockdown lines, leaf angles were larger than wild type, but the whole plant did not exhibit prostrate growth habit with the other mutations (Dong et al., 2013; Howard et al., 2014; Zhang et al., 2014). Over expression of *ZmCLA4* in rice led to narrower leaf angles than wild type rice.

In our *MdLazy1*-S:pGWB412 transgenic lines, we observed both leaf angle and branch angle differences between overexpression lines, wild type, knockdown and RNAi lines. In year one, *MdLazy1*-S:pGWB412 lines 1 and 7 under expressing *MdLazy1*-S showed large leaf angles, while line 6 overexpressing *MdLazy1*-S resulted in narrow leaf angles when compared to the non-transgenic control. In year two, we observed wide branch angles in the under expressing lines (line 1: 87.4°, line 7: 82.8°) and narrow branch angles in the over expressing line and 'Royal Gala' strain GL3 control (66.3° and 54.7° respectively)**.** Branch tip orientation in *MdLazy1* under expression lines was larger (line 1: 91.7°, line 7 80.7°) than in both WT and over expression lines (58° and 47.7° respectively). A tip orientation of 90° means that the shoot tip is

growing parallel to the ground, while a shoot tip greater than 90° indicates the shoot tip is growing in a downward direction. In 2018 branch angle and shoot tip orientation measurements were taken from six 'Cheal's Weeping', 'Red Jade' and 'Golden Delicious' trees, with at least 3 primary branches measured per tree. The average branch angles were 91.4°, 82.3° and 64.9° and the tip orientations were 126.8°, 134.7° and 29.2° for 'Cheal's Weeping', 'Red Jade' and 'Golden Delicious respectively (unpublished data). These data support that lines 1 and 7 under expressing *MdLazy1*-S are similar to weeping growth habit and line 6 is more similar to 'Golden Delicious' standard growth habit.

In contrast, all the *MdLazy1*-W:pGWB412 lines over expressed *MdLazy1*-W, resulting in wider leaf angles (104.6°-120.7° in lines, 58° in WT). Additionally the *Lazy1*-RNAi line had downward leaves. Taking into account the leaf angles observed in the *MdLazy1*-S:pGWB412 lines at a young age **(Figure S3.7)** and the branching angles/tip orientations observed at an older age, we expect *MdLazy1*-W:pGWB412 and RNAi lines to have similar growth. Preliminary branching data from the *MdLazy1*-W trees supports this notion (**Figure S3.9).** This strongly suggests that the $CT_{584}T>CC_{584}T$ mutation in *MdLazy1* is likely causal for the weeping phenotype in apple. As mentioned above, overexpression of Lazy1 in popular led to narrow branches and over expression in rice led to narrow leaf angles which is in agreement with our *MdLazy1*-S overexpression line, however over expression of the *MdLazy*-W in apple led to wider leaf angles which was observed in loss of function plants.

Protein *MdLazy1*-S was predicted to have a transmembrane domain from amino acids 183-200, while *MdLazy1*-W was not. The amino acid substitution of L195P in *MdLazy1*-W is within the predicted transmembrane region and could structurally change the protein form and therefore function at the plasma membrane. Prolines unique ring structure is less adaptable to α-

helix conformations and can cause kinks and improper folding (Betts & Russell, 2003). In an *Arabidopsis* study of *AtLazy1*, Sasaki and Yamamoto did a membrane association experiment and found *AtLazy1* fractionated in an insoluble fraction that contained membranous compartments, but was solubilized suggesting it is a peripheral membrane protein(Sasaki & Yamamoto, 2015). They further show the C-terminal end of *AtLazy1* interacted with microtubules in vitro (Sasaki & Yamamoto, 2015). It has been suggested that microtubules are involved in organ bending, auxin response and gravotropism (Bisgrove, 2008).

***MdLazy1* interactors**

Eleven *MdLazy1*-S protein interactors in yeast were identified. Five unique interactors were members of the regulator of chromatin condensation 1 (RCC1) family, three were members of the Brevis Radix family and the remaining three were *NAD kinase*, *adenosine kinase 2* and DNase-I super family proteins. The regulator of chromosome condensation 1 interactors all contained FYVE zinc finger domains. The RCC1 family and brevis radix proteins all contained the BRX domain (PF08381). Briggs et al. found that the BRX domain is also present in the PRAF (PH, RCC1, and FYVE)-like family proteins (Briggs, Mouchel, & Hardtke, 2006). Prosite scan of each RCC1 and brevis radix protein revealed they all contained the BRX domain in the C terminal end (de Castro et al., 2006) (**Figure S3.10)**. This domain may directly interact with *MdLazy1*. A recent study in rice identified *OsBrxL4*, a brevis radix like 4 gene that directly interacts with *Lazy1* at the plasma membrane (Z. Li et al., 2019). A study in *A. Thaliana* demonstrated that BRX protein is plasma membrane associated, but translocated to the nucleus upon auxin treatment (Scacchi et al., 2009). *Lazy1* in maize, interacted with GRMZM2G147243 (*IAA17*) in the nucleus and GRMZM2G026203 (*PKC*) at the plasma membrane (Dong et al., 2013). To confirm the interactors in planta, bimolecular fluorescence complementation (BiFC)

and western blots still need to be completed. All *MdLazy1* interactors were discovered with the standard version of the gene, therefore it would be interesting to see if the *MdLazy1*-W protein have the same interactors, especially those at the plasma membrane.

Conclusion

Using a 'Cheal's Weeping' OP population, the *W* locus previously identified was narrowed down to a 628Kb region. Within the region *MdLazy1* (MD13G1122400) was identified as a strong candidate gene for *W,* which is an orthologue of *Lazy1* genes involved in shoot gravotropism in rice, corn and Arabidopsis to be. Two alleles *MdLazy1*-S and *MdLazy1*-W were identified. *MdLazy1*-W has a non-synonymous L195P mutation that was exclusively present in all weeping trees screened. Transgenic trees overexpressing *MdLazy1*-W and those under expressing *MdLazy1*-S have similar phenotypes, suggesting the L195P mutation is likely causal for the weeping growth habit it apple.

Acknowledgements

This work was financially supported by a grant award (IOS-1339211) from NSF-Plant Genome Research Program. This work is co-authored by Raksha Singh, Ping Wang, Desen Zheng, Ewa Borejsza-Wysocka, Susan Brown and Kenong Xu.

REFERENCES

Arent, S., Pye, V. E., & Henriksen, A. (2008). Structure and function of plant acyl-CoA oxidases. *Plant Physiol Biochem, 46*(3), 292-301.

Betts, M. J., & Russell, R. B. (2003). Amino acid properties and consequences of substitutions. *Bioinformatics for geneticists, 317*, 289.

Bisgrove, S. R. (2008). The roles of microtubules in tropisms. *Plant Science, 175*(6), 747-755.

Borejsza-Wysocka, E. E., Norelli, J. L., Aldwinckle, H. S., & Ko, K. (1999). *TRANSFORMATION OF AUTHENTIC M.26 APPLE ROOTSTOCK FOR ENHANCED RESISTANCE TO FIRE BLIGHT*.

Briggs, G. C., Mouchel, C. F., & Hardtke, C. S. (2006). Characterization of the plant-specific BREVIS RADIX gene family reveals limited genetic redundancy despite high sequence conservation. *Plant physiology*, *140*(4), 1306-1316.

Chen, R., Rosen, E., & Masson, P. H. (1999). Gravitropism in Higher Plants. *Plant physiology, 120*(2), 343-350.

Chung, H. S., Koo, A. J. K., Gao, X., Jayanty, S., Thines, B., Jones, A. D., & Howe, G. A. (2008). Regulation and Function of Arabidopsis JASMONATE ZIM-Domain Genes in Response to Wounding and Herbivory. *Plant physiology*, *146*(3), 952-964.

Curtis, M. D., & Grossniklaus, U. (2003). A Gateway Cloning Vector Set for High-Throughput Functional Analysis of Genes in Planta. *Plant physiology*, *133*(2), 462-469.

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., & Velasco, R. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*.

Dardick, C., Callahan, A., Horn, R., Ruiz, K. B., Zhebentyayeva, T., Hollender, C., Whitaker, M., Abbott, A., & Scorza, R. (2013). PpeTAC1 promotes the horizontal growth of branches in peach trees and is a member of a functionally conserved gene family found in diverse plants species. *The Plant Journal, 75*(4), 618-630.

de Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., & Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research, 34*(suppl_2), W362-W365.

Digby, J., & Firn, R. D. (1995). The gravitropic set-point angle (GSA): the identification of an important developmentally controlled variable governing plant architecture. *Plant Cell Environ, 18*(12), 1434-1440.

Dong, Z., Jiang, C., Chen, X., Zhang, T., Ding, L., Song, W., Luo, H., Lai, J., Chen, H., Liu, R., Zhang, X., & Jin, W. (2013). Maize LAZY1 mediates shoot gravitropism and inflorescence development through regulating auxin transport, auxin signaling, and light response. *Plant Physiol, 163*(3), 1306-1322.

Dougherty, L., Singh, R., Brown, S., Dardick, C., & Xu, K. (2018). Exploring DNA variant segregation types in pooled genome sequencing enables effective mapping of weeping trait in Malus. *J Exp Bot*.

Doukhanina, E. V., Chen, S., van der Zalm, E., Godzik, A., Reed, J., & Dickman, M. B. (2006). Identification and functional characterization of the BAG protein family in Arabidopsis thaliana. *Journal of Biological Chemistry, 281*(27), 18793-18801.

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem bull, 19*, 11-15.

Earley, K. W., Haag, J. R., Pontes, O., Opper, K., Juehne, T., Song, K., & Pikaard, C. S. (2006). Gateway-compatible vectors for plant functional genomics and proteomics. *The Plant Journal, 45*(4), 616-629.

Friml, J., Wiśniewska, J., Benková, E., Mendgen, K., & Palme, K. (2002). Lateral relocation of auxin efflux regulator PIN3 mediates tropism in Arabidopsis. *Nature, 415*(6873), 806-809.

Godbole, R., Takahashi, H., & Hertel, R. (1999). The lazy mutation in rice affects a step between statoliths and gravity-induced lateral auxin transport. *Plant Biology, 1*(04), 379-381.

Hart, J. W. (1990). *Plant tropisms: and other growth movements*: Springer Science & Business Media.

Helliwell, C., & Waterhouse, P. (2003). Constructs and methods for high-throughput gene silencing in plants. *Methods, 30*(4), 289-295.

Helliwell, C. A., Wesley, S. V., Wielopolska, A. J., & Waterhouse, P. M. (2002). High-throughput vectors for efficient gene silencing in plants. *Functional Plant Biology, 29*(10), 1217-1225.

Hill, J. L., & Hollender, C. A. (2019). Branching out: new insights into the genetic regulation of shoot architecture in trees. *Current Opinion in Plant Biology, 47*, 73-80.

Hofmann, K., & Stoffel, W. (1993). TMbase - A database of membrane spanning proteins segments. *Biological Chemistry Hoppe-Seyler, 374*(166).

Howard, T. P., III, Hayward, A. P., Tordillos, A., Fragoso, C., Moreno, M. A., Tohme, J., Kausch, A. P., Mottinger, J. P., & Dellaporta, S. L. (2014). Identification of the Maize Gravitropism Gene lazy plant1 by a Transposon-Tagging Genome Resequencing Strategy. *PLoS One, 9*(1), e87053.

Just, B. J. (2001). *Molecular markers for weeping plant habit and powdery mildew (Podosphaera Leucotricha) resistance from the ornamental crabapple 'Red Jade'*: Cornell University, August.

Li, P., Wang, Y., Qian, Q., Fu, Z., Wang, M., Zeng, D., Li, B., Wang, X., & Li, J. (2007). LAZY1 controls rice shoot gravitropism through regulating polar auxin transport. *Cell Research, 17*, 402.

Li, Z., Liang, Y., Yuan, Y., Wang, L., Meng, X., Xiong, G., Zhou, J., Cai, Y., Han, N., Hua, L., Liu, G., Li, J., & Wang, Y. (2019). OsBRXL4 Regulates Shoot Gravitropism and Rice Tiller Angle through Affecting LAZY1 Nuclear Localization. *Molecular Plant, 12*(8), 1143-1156.

Meisel, L., Fonseca, B., Gonzalez, S., Baeza-Yates, R., Cambiazo, V., Campos, R., Gonzalez, M., Orellana, A., Retamales, J., & Silva, H. (2005). A rapid and efficient method for purifying high quality total RNA from peaches (Prunus persica) for functional genomics analyses. *Biol Res, 38*(1), 83-88.

Mika, A. (1992). *TRENDS IN FRUIT TREE TRAINING AND PRUNING SYSTEMS IN EUROPE*.

Moore, I. (2002). Gravitropism: Lateral Thinking in Auxin Transport. *Current Biology, 12*(13), R452-R454.

Morita, M. T. (2010). Directional Gravity Sensing in Gravitropism. *Annual Review of Plant Biology, 61*(1), 705-720.

Nelson, B. K., Cai, X., & Nebenfuhr, A. (2007). A multicolored set of in vivo organelle markers for co-localization studies in Arabidopsis and other plants. *Plant J, 51*(6), 1126-1136.

Roychoudhry, S., Del Bianco, M., Kieffer, M., & Kepinski, S. (2013). Auxin Controls Gravitropic Setpoint Angle in Higher Plant Lateral Branches. *Current Biology, 23*(15), 1497-1504.

Sampson, D., & Cameron, D. (1965). *INHERITANCE OF BRONZE FOLIAGE EXTRA PETALS AND PENDULOUS HABIT IN ORNAMENTAL CRABAPPLES.* Paper presented at the Proceedings of the American Society for Horticultural Science.

Sang, D., Chen, D., Liu, G., Liang, Y., Huang, L., Meng, X., Chu, J., Sun, X., Dong, G., Yuan, Y., Qian, Q., Li, J., & Wang, Y. (2014). Strigolactones regulate rice tiller angle by attenuating shoot gravitropism through inhibiting auxin biosynthesis. *Proceedings of the National Academy of Sciences, 111*(30), 11199-11204.

Sasaki, S., & Yamamoto, K. T. (2015). Arabidopsis LAZY1 is a peripheral membrane protein of which the carboxy-terminal fragment potentially interacts with microtubules. *Plant Biotechnology, 32*(1), 103-108.

Scacchi, E., Osmont, K. S., Beuchat, J., Salinas, P., Navarrete-Gómez, M., Trigueros, M., Ferrándiz, C., & Hardtke, C. S. (2009). Dynamic, auxin-responsive plasma membrane-to-nucleus movement of Arabidopsis BRX. *Development, 136*(12), 2059-2067.

Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D. M., Singh, K. B., Dodds, P. N., & Taylor, J. M. (2017). LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Scientific Reports, 7*, 44598-44598.

Taniguchi, M., Furutani, M., Nishimura, T., Nakamura, M., Fushita, T., Iijima, K., Baba, K., Tanaka, H., Toyota, M., Tasaka, M., & Morita, M. T. (2017). The Arabidopsis LAZY1 Family Plays a Key Role in Gravity Signaling within Statocytes and in Branch Angle Control of Roots and Shoots. *The Plant Cell, 29*(8), 1984-1999.

Toyota, M., & Gilroy, S. (2013). Gravitropism and mechanical signaling in plants. *American journal of botany, 100*(1), 111-125.

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Research, 46*(5), 2699-2699.

Vandenbrink, J. P., & Kiss, J. Z. (2019). Plant responses to gravity. *Seminars in Cell & Developmental Biology, 92*, 122-125.

Wang, L., Cai, W., Du, C., Fu, Y., Xie, X., & Zhu, Y. (2018). The isolation of the IGT family genes in Malus × domestica and their expressions in four idiotype apple cultivars. *Tree Genetics & Genomes, 14*(4), 46.

Xu, D., Qi, X., Li, J., Han, X., Wang, J., Jiang, Y., Tian, Y., & Wang, Y. (2017). PzTAC and PzLAZY from a narrow-crown poplar contribute to regulation of branch angles. *Plant Physiology and Biochemistry, 118*, 571-578.

Yoshihara, T., & Iino, M. (2007). Identification of the Gravitropism-Related Rice Gene LAZY1 and Elucidation of LAZY1-Dependent and -Independent Gravity Signaling Pathways. *Plant and Cell Physiology, 48*(5), 678-688.

Yoshihara, T., & Spalding, E. P. (2017). LAZY Genes Mediate the Effects of Gravity on Auxin Gradients and Plant Architecture. *Plant physiology, 175*(2), 959-969.

Yoshihara, T., Spalding, E. P., & Iino, M. (2013). AtLAZY1 is a signaling component required for gravitropism of the Arabidopsis thaliana inflorescence. *The Plant Journal, 74*(2), 267-279.

Zhang, J., Ku, L. X., Han, Z. P., Guo, S. L., Liu, H. J., Zhang, Z. Z., Cao, L. R., Cui, X. J., & Chen, Y. H. (2014). The ZmCLA4 gene in the qLA4-1 QTL controls leaf angle in maize (Zea mays L.). *Journal of Experimental Botany, 65*(17), 5063-5076.

Zhang, N., Yu, H., Yu, H., Cai, Y., Huang, L., Xu, C., Xiong, G., Meng, X., Wang, J., Chen, H., Liu, G., Jing, Y., Yuan, Y., Liang, Y., Li, S., Smith, S. M., Li, J., & Wang, Y. (2018). A Core Regulatory Pathway Controlling Rice Tiller Angle Mediated by the LAZY1-Dependent Asymmetric Distribution of Auxin. *The Plant Cell, 30*(7), 1461-1475.

Zhang, X., Henriques, R., Lin, S.-S., Niu, Q.-W., & Chua, N.-H. (2006). Agrobacterium-mediated transformation of Arabidopsis thaliana using the floral dip method. *Nature Protocols, 1*, 641.

Supplementary Figures

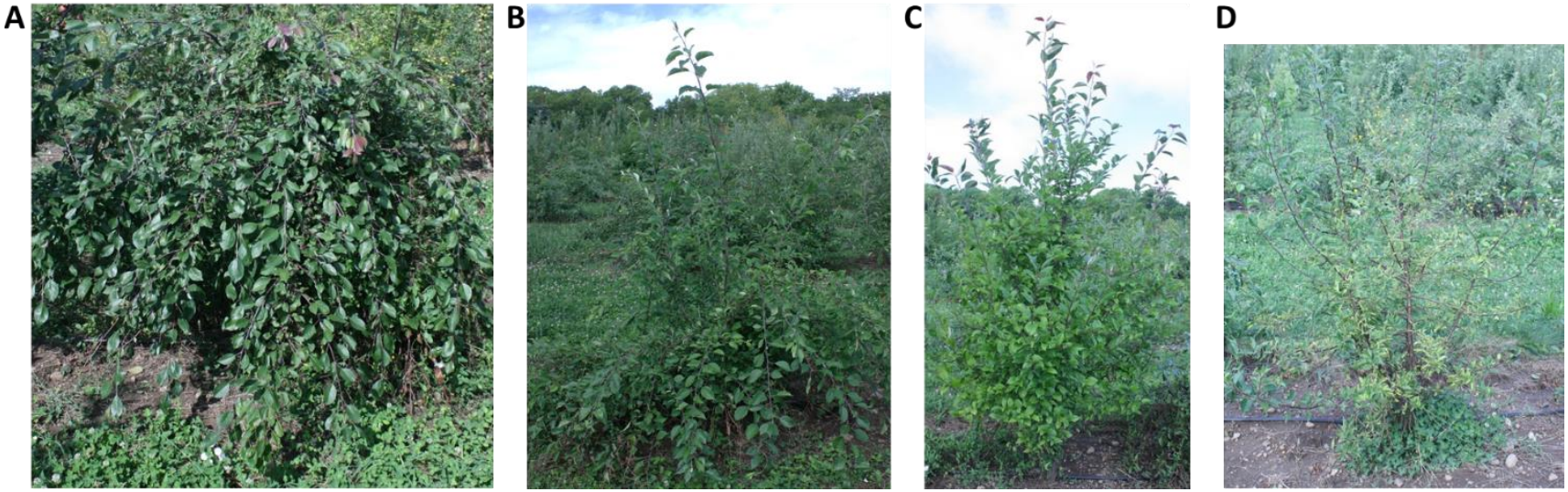


Figure S3.1. ‘Cheal’s Weeping’ OP progeny. Examples of A. weeping, B. weeping-like, C. standard and D. standard-like phenotypes

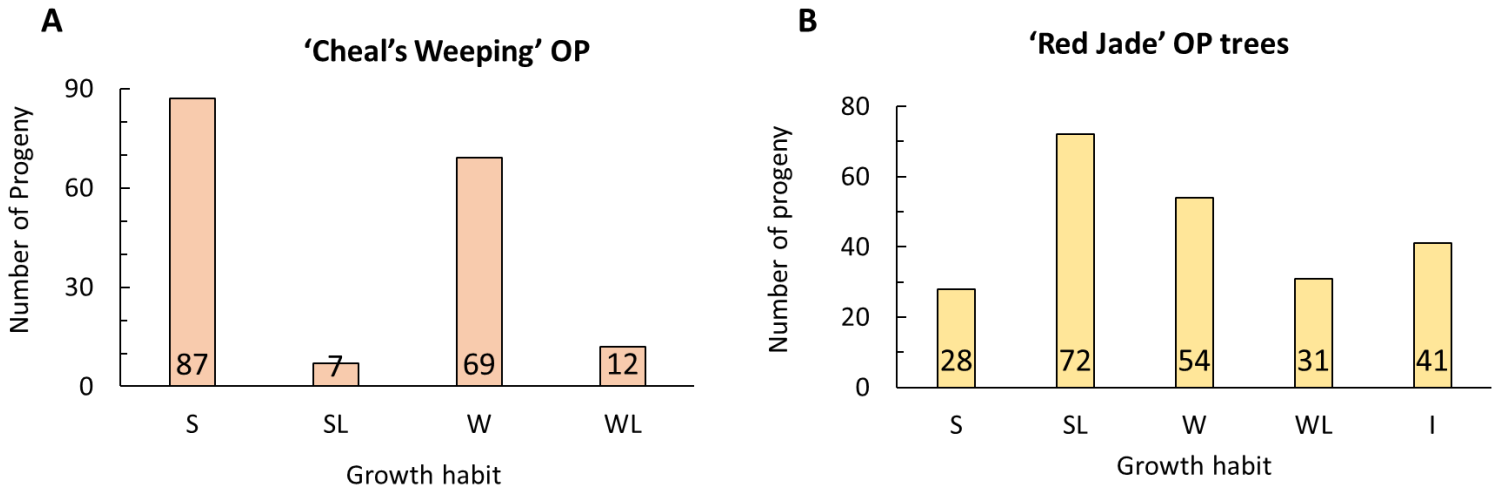


Figure S3.2. Phenotypic evaluation of growth habit in populations: **A.** 'Cheal's Weeping' OP and **B.** 'Red Jade' OP Growth habit phenotypes are S (standard), SL (Standard-like), W (weeping) , WL (weeping like) and I (intermediate respectably. Chi-square tests showed that the segregation of weeping (-like) and standard (-like) growth habits fit the 1:1 ratio the population ($\chi 2$=2.719, P=0.099). Number of trees for each category is shown in yellow.

MdLazy1-W cDNA sequence

ATGAAGTTACTAGGTTGGATGCATCGTAAGTTTCGGCAGAATAGCAACGATCCATTTAAAGTTTTCGTCGTTGGGCAGCCATCGCTCGACGATCAACAATACTATCCGAAGCCAAACTGCGGCACGAAACCC
TTTAAACAAGCCCCGAGAGATCAGCACGTTCGAAAATCTTTCAACGGTCTAGAAGCAGCTAGGACAGAAGAAGAATACTACGAAGAAGAATCATCTGCTGAAGCATCTGAATACTTCTATGGATTTCTTGC
AATCGGTACTCTTGGCAGCTCGGACCAAGTGACCACTGAACCATCAACTCCAACGCTTGGAATCTCTGTGGAGAACATAACCGAAAAAGAAACTGAGGCCACGGAGAACGAGCTGAAGCTCATCAATGAT
GAATTGGAGAGAGAGTTGGTGGCCGATCCGGCGAAGGATGATATTTGCAATGATTCATCTGGAAGAAACAGCTATGTCAGCAATGGAAGAAGTAGCCATGGAAGCACCATTACGCTTAGTGGGAAGGCG
CTGGAAGGCGCAGAGAGCAACGGGATCAATGGAACTGCAGTGTGCCCGCTCCAGGGATATCCTTTCGGGTCGGCATATGAATTGTCTGAAACAACAAGCATGACAAAGAAGGAACACAGGACATCCCTT
GGTGAGCTATTTCAGAGGACTAAATTGGCAGAGGAGATTTCTGGAGCAAAATTAAACAAGGAGGAGAAGCGAGCAGAGAAGGAAGCGGATAAGTCCGCCATGCACTTGATGAAAAAGATGCTCAAGAA
AAAAATGATTCACGCTTCTTCTCGTGGCTCCGGCGGAGCTGCTGATCACGCTTCAGCAGAAACAAAACTCAATAAGATCCTTCACATGTTTCATAGAAAAGTTCACCCTGAAACCTCGGCGGCTCAGCAAAA
ACCTGGTAAGTACGAGAGGAACCAAACCAAGAAGAAAACAAGCAGTGAGGGGACTTACAGCAATGGAGATCAGGAGGATGATATCATCTTATATCCTCAGCAAGGATTTTCTTCAAATCCGAGCATGCGG
CGCTACAAGAGCCAATCAAACCCGCCCCAAATCGCGCTTAGCGGCATGGATTCAAATGATAACAGGGAGCACTGGATCAAAACAGATGCAGACTACCTAGTCTTGGAGCTGTAG

MdLazy1-S cDNA sequence

ATGAAGTTACTAGGTTGGATGCATCGTAAGTTTCGGCAGAATAGCAACGATCCATTTAAAGTTTTCGTCGTTGGGCAGCCATCGCTCGACGATCAACAATACTATCCGAAGCCAAACTGCGGCACGAAACCC
TTTAAACAAGCCCCGAGAGATCAGCACGTTCGAAAATCTTTCAACGGTCTAGAAGCAGCTAGGACAGAAGAAGAATACTACGAAGAAGAATCATCTGCTGAAGCATCTGAATACTTCCATGGATTTCTTGCA
ATCGGTACTCTTGGCAGCTCGGACCAAGTGACCACTGAACCATCAACTCCAACGCTTGGAATCTCTGTGGAGAACATAACCGAAAAAGAAACCGAGGCCACGGAGAACGAGCTGAAGCTCATCAATGATGA
ATTGGAGAGAGAGTTGGTGGCTGATCCGGCGAAGGATGATATTTGCAATGATTCATCTGGAAGAAACAGCTATGTCAGCAATGGAAGAAGTAGCCATGGAAGCACCATTACGCTTAGTGGGAAGGCGCT
GGAAGGCGGAGAGAGCAACGGGATCAATGGAACTGCAGTGTGCCCGCTCCAGGGATATCTTTTCGGGTCGGCATATGAATTGTCTGAAACAACAAGCGTGACAAAGAAGGAACACAGGACATCCCTTGG
TGAGCTATTTCAGAGGACTAAATTGGCAGAGGAGATTTCTGGAGCAAAATTAAGCAAGGAGGAGAAGCGAGCAGAGAAGGAAGCGGATAAGTCCGCCATGCACTTGATGAAAAAGATGCTCAAGAAAAA
AATGATTCACGCTTCTTCTCGTGGCTCCGGCGGAGCTGCTGATCACGCTTCAGCAGAAACAAAACTCAATAAGATCCTTCACATGTTTCATAGAAAAGTTCACCCTGAAACCTCGGCGGCTCAGCAAAAACCT
GGTAAGTACGAAAGGAACCAAACCAAGAAGAAAACCAGCAGTGAGGGGACTTACAGCAATGGAGATCAGGAGGATGATATCATCTTATATCCTCAGCAAGGATTTTCTTCAAATCCGAGCATGCGGCGCT
ACAAGAGCCAATCAAACCCGCCCCAAATCGCGCTTAGCGGCATGGATTCAAATGATAACAGGGAGCACTGGATCAAAACGGATGCTGACTACCTAGTCTTGGAGCTGTAG

Figure S3.3. Coding sequences for the weeping and standard alleles of *MdLazy1* found in 'Cheal's Weeping'. The 584bp SNP is highlighted in red. The stop codon is underlined.

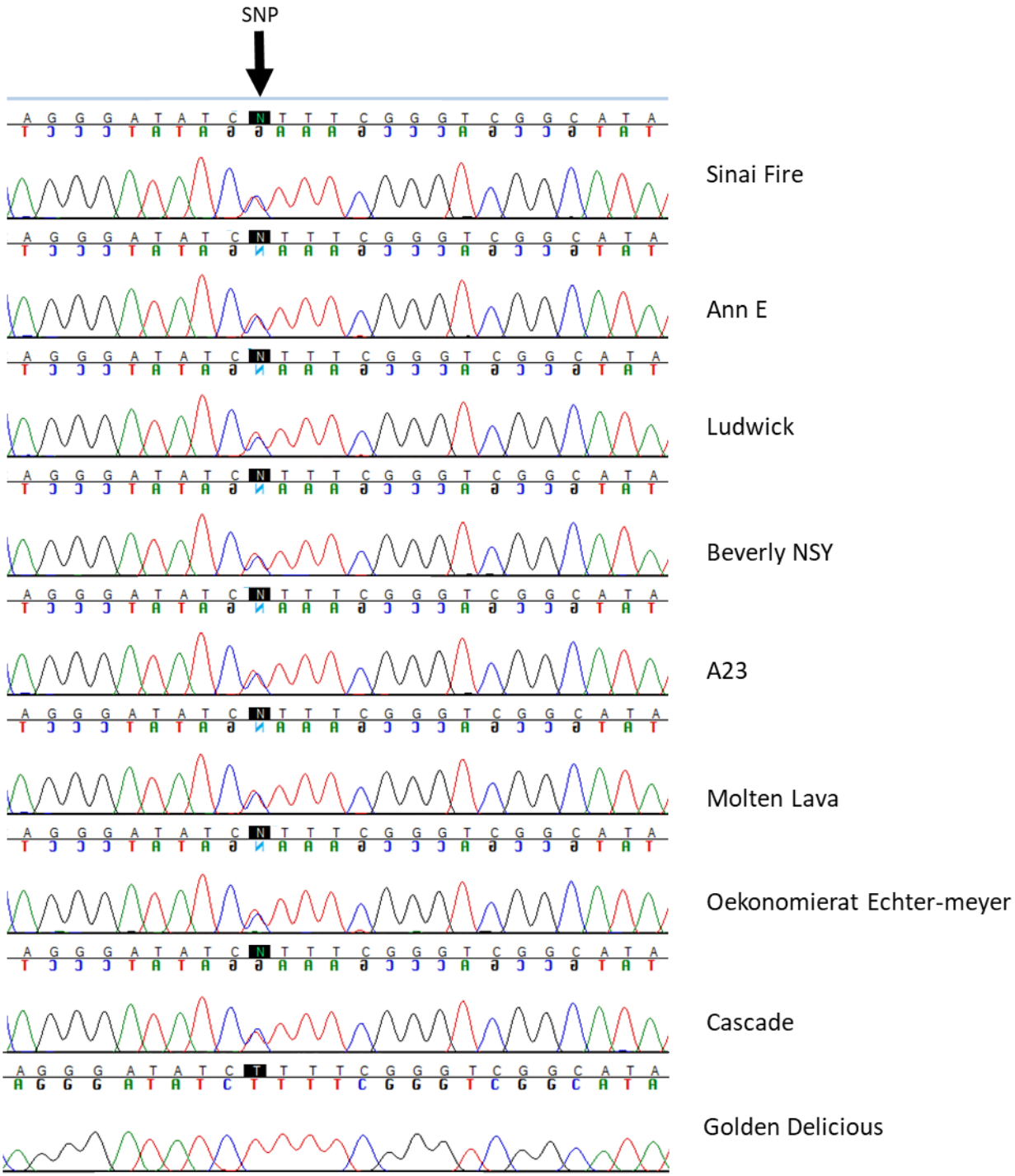


Figure S3.4: Confirmation of C584 (*MdLazy1*-W) SNP in diverse weeping varieties. DNA was PCR amplified and used for sequencing. The target SNP is highlighted in black. Weeping varieties have an 'N' at the position 584 meaning they are heterozygous. Blue line represent cytosine while red, black and green lines represent thymine, guanine and adenosine respectively.

GTGTGTGGAGTCATGTATAACGAATGAGGATTTGGGGATTTTCTTCATGTCCATTTATCGTAATCATACGGTTATAAATTATTTTAAATTTTTTATTTAATTGAATACAAAT
AATACCTGATGAAAATTGACCTTACGATCAGTGGCGAAGCCACAACAGGGCAAGGCGAGGCGGCTGCCCCTCCGGTCATCGGAAATTCACATGCAGAGCAAGGAATC
CACCCCTTCGATAGTCGAAAAGCCACATAAAGAGCCATGGTGTCGGTGGGCTGCTTGTTGCAGAGGAGGGCAACTCCTACGCGTGGAAGAAGAACAAAAAAGAAAAG
TTGTTTTAAAAATATCTTGAACCGGACGACGTCGTTTCAGCATTCCTTTTAAATGAAACTTTGCGTTTTGTTACTTTTTTTTTTCTATCCTTCTTTTGTTGCTTTTCACCGTTTT
TCCTTTCCCCTTTTTATTTTTTTTCTTTTTCCTTCTATCTTTTTTCCATTTGACTTGGTTCCTTAGTTTTTGGTTAAATGGTTTGGTTTGGTCTCTCTTTCTTTTGATTTGATCTTC
TGATTGGTCCCTAATTATACTTTAAAAAACATAAGTGCTTCAATGAAATATTTTAGATAAACCATACCCGCAAAATACCACATAAAATAAGACTTTCAATACTAAAATTACC
TTAATGTTGTTAATACAATATGGAGGAGGTTATTTCATACTCAAATTTGTATGTTTTTTTTTCTAGCCATCACTACTTTCAATGAATTTCTGATGTGATTTTGAGGACTTGTT
TCATGTTTCAAATAATATGAAAATTAATGGAAAGGGTTTGAAAAATTTTGAGTTTTAACCAAAAGAACAAAAATGTGTTGTAAGTGAATAGTACTAAGAGTGACTTTTTA
GAGTAAAAATGTCTTTAACATTAAAAGTGAACAATACCATGAGTGTTTCGTTAAGACTCCCTAATTTAAATCCATCTGGATTCCTGCCTAGATCTCGCCTAGGCACCTAGA
TGCTAGGTTCTAGTCCATTGTTTGACTAGCGCTTAACATTTTTTAGGATCTAATCCTAGCCAATTTAAACTTCTTACATTGGCTTTGATATTTTTGCTTACAAATGCTTTGGT
GGGGAGACTATTTTCTGCTATAAATATTG:ACTCAAAACTTTGCACTCTCCAATGTGAAAGCAAACGCTTCGATTGAACCTTACACTCTTTTGAATTTCGCCCCTCCCCTCG
ACAAATCTTAGCTTCACCACTGCTTACGATATACGATGAATGAATACGATTAAAGAATCTTCTGGATCCTCATTCGTGTATTACTGTACAAAAAGATATAAAGTGGTTGTC
CCTGAATTATTCACCACAAATTGATACTATGAAATATAAACCCTACAAATTTCACACAATAATATATATATAAACCCTACAATCAATCTCACATTATAAACGTATTTGGGGT
CGGGACCCACTTGCATCCATTGGCCTCTTGATTTGATCTTGACAAAGGATAGCCAGCCCAGGTTTGTAACTATCTGTTTGTACTTCTCTGCCCTCATTCTCTTCAGAGGCT
GGAGGCTTTCGGGTAAGCATTTGACACCTTATTCATAAATAAGCATAAAGACTGACTAATAATGAAAAGAAGTTACACTTTCAGCCTCTCCTAGCAAAGCTCAAGTCATC
TAAAAGAAAGGTTACAAG

Figure S3.5. *MdLazy1*-W promoter sequence used for the ß-glucuronidase assay. The promoter is 1.69kb long and ends immediately before the ATG start codon. The sequence was put into reporter vector pMDC164C.

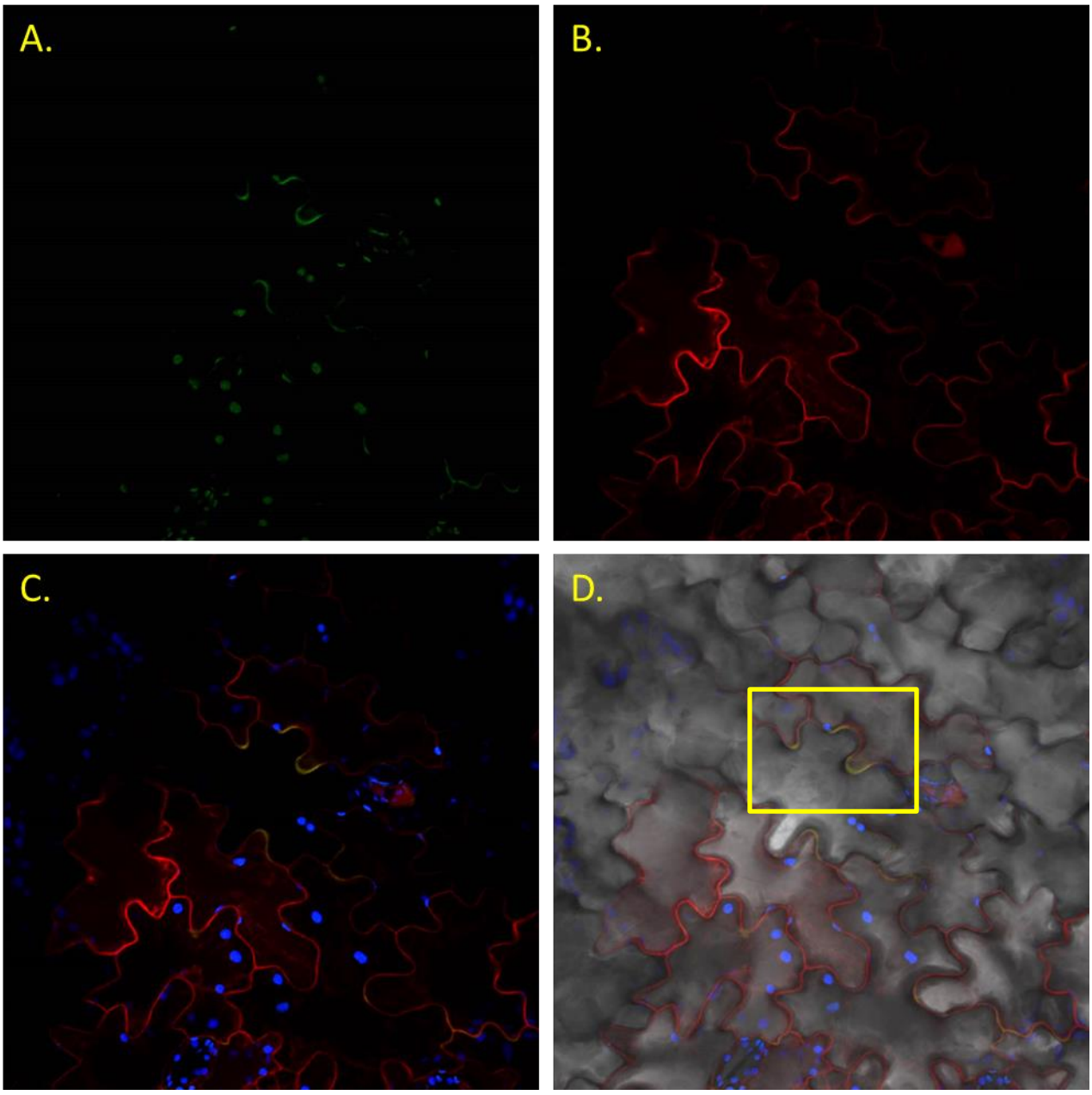


Figure S3.6. Co-localization of *MdLazy1*-S to plasma membrane. **A.** *Lazy1-S*:GFP localization with 488nm laser. **B**. Plasma membrane-mCherry marker. **C.** Overlay of *Lazy-S*:GFP, Plasma membrane-mCherry and chlorophyll (shown in blue). **D.** Overlay with bright field. Yellow lines highlighted in the box show *MdLazy1*-S:GFP and plasma membrane marker overlapping confirming that *MdLazy1*-S localizes to the plasma membrane.

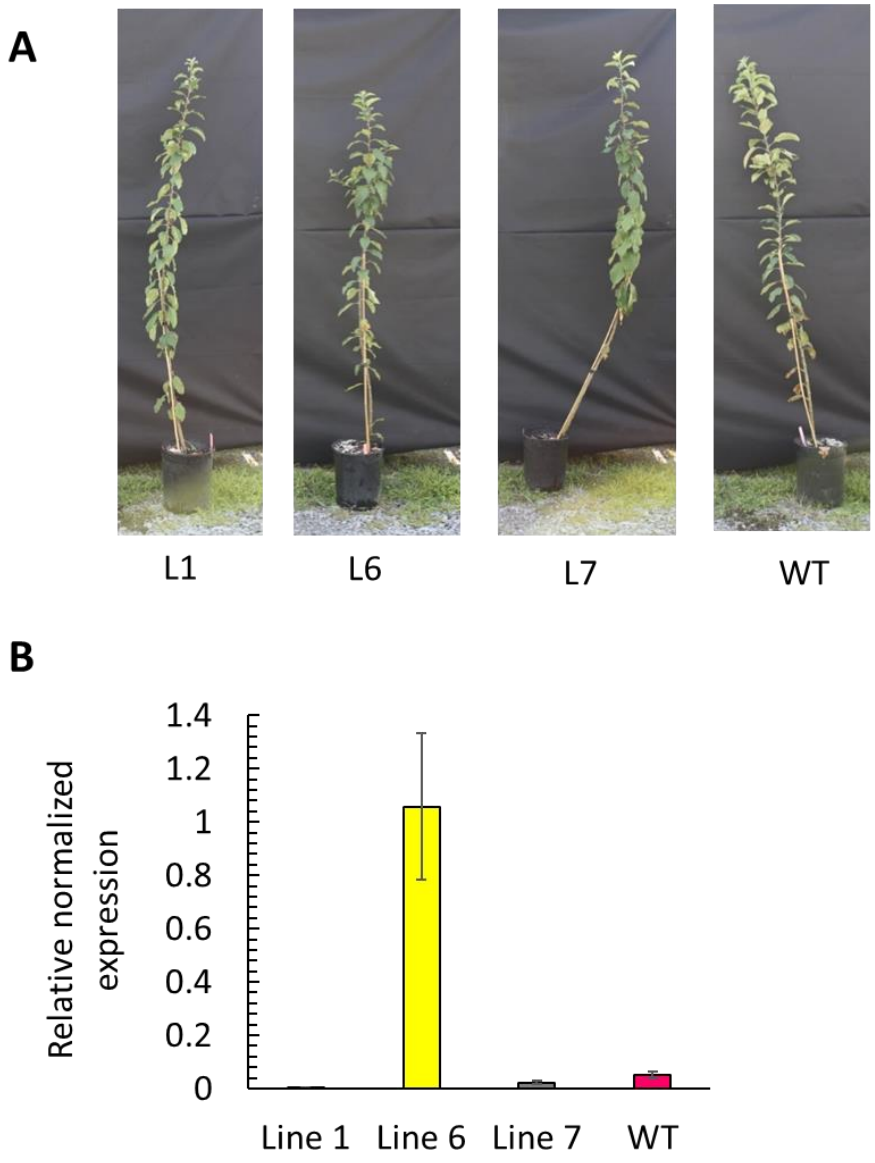


Figure S3.7. Reduced expression of *MdLazy1*-S in leaf petioles leads to drooping leaves: **A.** Lines 1, 4, 7 and 10 have drooping like leaves. WT and line 6 have horizontal leaves. **B.** Expression of *MdLazy1* in plant petioles. Lines 1, 4, 7 and 10 have lower expression of *MdLazy1* compared to WT while line 6 is overexpressed.

Figure S3.8. *MdLazy1*-S:pGWB412 and *MdLazy1*-W:pGWB412 transgenic plants: At a young age, leaf angles were visibly different between A. *MdLazy1*-S under expression line 7, **B.** *MdLazy1*-S over expression line 6, **C.** Royal Gala stain control, **D.** *MdLazy1*-W line 4 **E.** 'Royal Gala' control and **F.** *MdLazy1*-RNAi. Tree were approximately 2 months old at time of photo.

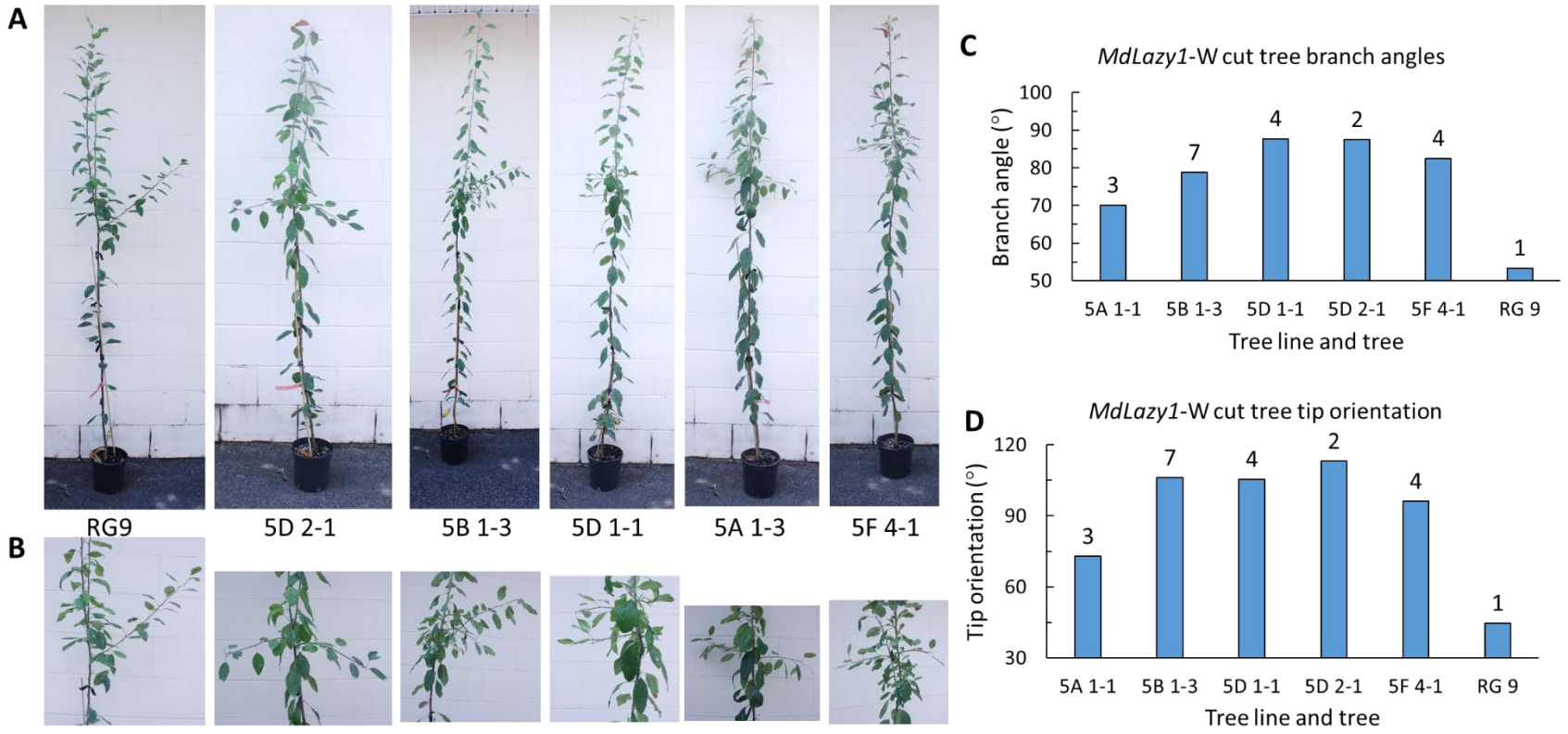


Figure S3.9. *MdLazy1*-W:pGWB412 transgenic plants: **A**. Apical shoot tips were cut to promote branching. These pictures were taken two months after cutting. Compared to the wild type (RG9) all the lines have wide branch angles and horizontal tip orientations. The same leaf angles are observed here as when the trees were very young in Figure 3.6 **B.** Close up of branches **C.** Average branch angle for trees. **D.** Average tip orientations. Number of branches are indicated per tree.

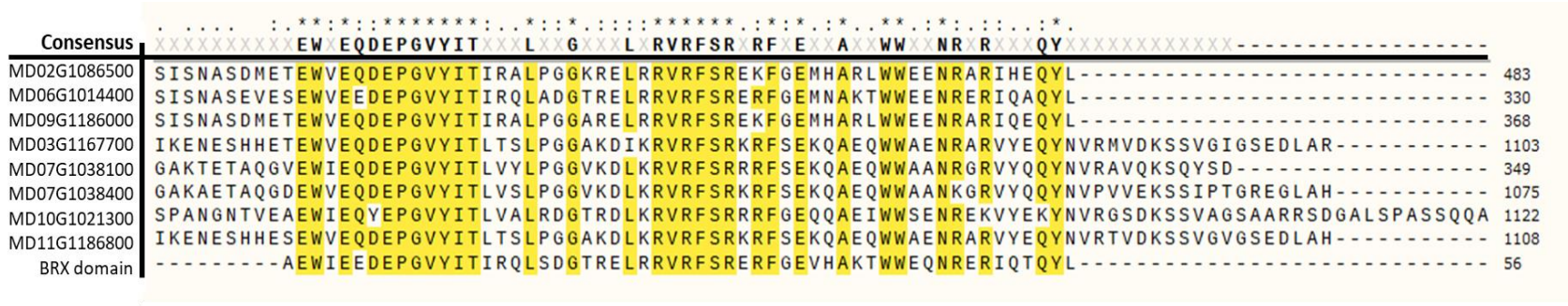


Figure S3.10. Protein alignment of *MdLazy1* Brevis radix and RCC1 interactors at the C-terminal end. All interactors share a BRX domain (shown in last row). The proteins did not show any other areas of similarities (not shown). Amino acids highlighted in yellow are highly conserved (AA is in 75% of proteins). Above the consensus sequence: '*' single and fully conserved residue, ':' strongly similar conservation (>0.5 on PAM250 matrix), ' .' weakly similar conservation (< 0.5 on PAM250 matrix).

Supplementary Tables

Table S3.1 List of primers used in study. All HRM markers were designed from 'Cheal's Weeping' SNPs. * indicated informative HRM markers

Lazy1 gene amplification	Forward Primer	Reverse Primer	Purpose	Gene target	Ch13 location (bp)	Notes
Lazy1F/ Lazy1R-NS	ATGAAGTTACTAGGTTGGATGC	CAGCTCCAAGACTAGGT	Amplify *MdLazy* cDNA sequence for PEG103 construct			No stop codon
Lazy1F/ Lazy1R-S	ATGAAGTTACTAGGTTGGATGC	TCACAGCTCCAAGACTAGGT	Amplify *MdLazy* cDNA sequence for pGWB412 construct and gDNA sequence			With stop codon
Lazy1Pro-F/ LAZY1pro R2	GTGTGTGGAGTCATGTATAACGAAT	CTTGTAACCTTTCTTTTAGATGACTTGA	Amplify promoter for pMDC164c Gus construct			
C13SR_8547F1 / MdLAZYPrF2	TCCTTTGTCAAGATCAAATCAAGA	GTCCCTGAATTATTCACCACAA	SSR marker for promoter deletion			
XM1688Ex1F/R	TCTTGGTTTGGTTCCTTTCG	GGACATCCCTTGGTGAGCTA	qPCR primers for *MdLazy1* expression analysis			MD13G1122400
Actin F/R	GGCTGGATTTGCTGGTGATG	TGCTCACTATGCCGTGCTCA	qPCR primers for Actin control expression analysis			EB136338, a Gala actin gene/EST used as a reference in qRT-PCR.
Lazy1RNAi F/R	GGGGACAAGTTTGTACAAAAAAGCAGGCTATGTAGCCATGGAAGCAGCATT	GGGACCACTTTGTACAAGAAAGCTGGGTATCTTGGTTTGGTTCCTTTCG	RNAi *MdLazy1* primers			AttB sites are unerlined
MD13G1120500F/R	GGCAAAGCGTCGAACGAATC	GAATAGATTCTCGCCGGGCA	HRM for mapping	MD13G1119000	8694878	Weeping is T/C, standard is T/T
MD13G1121500F/R	ATTACAGAGCTGGGGATGCG	CGTTCTTCATTGTCGGACGG	HRM for mapping	MD13G1120	8858100	Weeping is A/G, standard is A/A

				500		
MD13G1121200F/R	GCCCGGAAAGCAGACTGTAT	AAAACACCTTGGCTGCAGTC	HRM for mapping	MD13G1121200	8898459	Weeping is T/C, standard is C/C
MD13G1119000F/R	TGACGTTTGCCTACCCTTTCT	GCTAAACGAAGCAAATAAGACCCT	HRM for mapping	MD13G1121500	8949908	Weeping is T/C, standard is C/C
lzySNP584HRMF/R	ACAATTCATATGCCGACCCGA	CAATGGAACTGCAGTGTGCC	HRM marker to detect Lazy1-W base 584 SNP	MD13G1122400	9031158	Weeping is C/T, standard is T/T
MD13G1122800F/R	CGTCTTCGTCGTCGTCATCA	GGCGTCCGTGTACAGATGAA	HRM for mapping	MD13G1122800	9100385	Weeping is A/G, standard is A/A
MD13G1123200F/R	CCCGTCCAACTCCAAAACCT	TCAGCCAATTTACACAAATTCTAACA	HRM for mapping	MD13G1123200	9133588	Weeping is C/T, standard is C/C
MD13G1124100F/R	CACGTCTCCCGCTGATAGTT	TTTGATGGGTGACACGTGGT	HRM for mapping	MD13G1124100	9209949	Weeping is C/T, standard is C/C
MD13G1125400F/R	GGGCTTCTAAACCTGTCCCC	AGGTCCAGAAAACATTGGGTGA	HRM for mapping	MD13G1125400	9300053	Weeping is G/A, standard is G/G
MD13G1127300F	CCATTGACGCCATGATCTTGG	CTGTTGTCTATGCCCGGGTT	HRM for mapping	MD13G1127300	9574150	Weeping is C/T, standard is C/C

CHAPTER 4

Exploring DNA variant segregation types enables mapping of columnar apple recessive repressors and a Co-guided network

Abstract

Columnar apples trees, originated from a bud mutation 'Wijcik McIntosh', develop a simple canopy and set fruit on spurs. These characteristics make them an important genetic resource for improvement of tree architecture. Genetic studies have uncovered that columnar growth habit is a dominant trait and is caused by a retroposon insertion that induces the expression of the neighboring gene *Co* encoding a 2OG-Fe(II) oxygenase. Here we report the genetic mapping of two loci of recessive repressors *c2* (on Chr10) and *c3* (on Chr9) that are linked to repression of the retroposon-induced *Co* gene expression and associated columnar phenotype in 275 F_1 progeny, which were developed from a reciprocal cross between two columnar selections heterozygous at the *Co* locus. The mapping was accomplished by sequencing a genomic pool comprising 18 columnar progenies and another pool of 16 standard progenies that also carry the retroposon insertion, and by exploring DNA variants of segregation types that are informative for mapping recessive traits in apple. The informative segregation types include <**h**k × **h**k>, <**l**m × **ll**>, <**nn** × **n**p>, <**l**m × mm>, and <pp × **n**p>, where each letter denotes one of the four DNA bases and the letters in bold represent variants in relation to the reference genome. The alleles in each first and third positions are assumed in linkage with the recessive repressors' allele in the two parents, respectively. Using RNA-seq analysis, we further revealed that the *Co* gene together with the differentially expressed genes under loci *c2* and *c3* form a co-expression gene-network module associated with growth habit, in which 12 MapMan Bins were enriched.

Introduction

In the mid-20th century, the successful genetic improvement in plant architecture had led to a drastic yield increase worldwide in field crops, particularly corn, rice and wheat. The landmark accomplishment in agriculture has been known as "the Green Revolution" (Evenson & Gollin, 2003). To keep apple trees in optimal shape for fruit production in orchards, horticulturists have been improving tree pruning and training systems and developing different dwarfing rootstocks (Robinson, Hoying, Sazo, DeMarree, & Dominguez, 2013). Although such efforts are effective for productivity improvement in modern orchards, apple production costs also have been increased markedly due to manual tree pruning and fruit harvest (Taylor & Granatstein, 2013; West, Sullivan, Seavert, & Castagnoli, 2012). There is a strong demand for automation of labor-intensive orchard tasks, especially fruit harvest. The complex and dynamic tree canopy and variable fruit bearing sites have been the major challenges to automating fruit harvest although motorized platforms that can improve fruit harvest efficiency are available and promising porotypes of robotic fruit harvesters are being tested.

Columnar apple trees, which originally were discovered as a bud mutation from 'McIntosh', called 'Wijcik McIntosh' in 1960s (Lapins, 1969), develop a canopy much simpler than standard apple trees do due to their reduced number of branches and vertically growing branches. Columnar cultivars usually set fruit on spurs from old woods such as the main trunk and primary limbs, requiring little pruning. These `characteristics make columnar architecture an ideal fit for automation of pruning and harvesting. To take advantage of these desirable characteristics, 'Wijcik McIntosh' has been used in many breeding programs to develop new and improved columnar apple cultivars (Moriya et al., 2009; Tobutt, 1984). However, a major issue of existing columnar apple cultivars is their strong tendency to biennial bearing (Lauri &

Lespinnasse, 1993; Otto, Petersen, Brauksiepe, Braun, & Schmidt, 2014; Petersen & Krost, 2013) while some studies observed that approximately 5% of columnar progeny show regular annual bearing in breeding populations, indicating columnar and biennial bearing are not always linked (Blazek, 2013; Vávra, BlaÅ¾ek, Vejl, & JonÃ¡Å¡, 2017).

Columnar growth habit has been a major subject in apple genetic studies. An early investigation reported that the columnar growth habit was controlled by a dominant gene, called *Co* (Lapins, 1976). The *Co* locus was mapped to linkage group 10 by many studies (Conner, Brown, & Weeden, 1997, 1998; Fernandez-Fernandez et al., 2008; Hemmat, Weeden, Conner, & Brown, 1997; Kenis & Keulemans, 2007; Kim, Song, Hwang, Shin, & Lee, 2003; Moriya et al., 2009; Tian, Wang, Zhang, James, & Dai, 2005; Zhu, Zhang, Li, & Wang, 2007), and was characterized in detail (T. Bai et al., 2012; Baldi et al., 2013; Morimoto & Banno, 2015; Moriya, Okada, Haji, Yamamoto, & Abe, 2012). Sequencing analyses of the *Co* locus revealed an 8.2-kb DNA insertion (a long terminal repeat retroposon) in an inter-genic region to be genetically causal for the columnar phenotype, as the insertion is not present in 'McIntosh' while in 'Wijcik McIntosh' (Okada et al., 2016; Otto et al., 2014; Wolters, Schouten, Velasco, Si-Ammour, & Baldi, 2013). Despite lacking direct interruption of any genes, the retroposon insertion increased the expression of a nearby gene encoding a 2-oxoglutarate (OG) and Fe(II)-dependent oxygenase in columnar (Okada et al., 2016; Otto et al., 2014; Wolters et al., 2013), which is called *Co* in this study. The expression of the *Co* gene was found specific to root in standard apples, suggesting its expression in shoot and leaves in columnar apples is ectopic (Wada et al., 2018). Moreover, transgenic apples over-expressing the *Co* gene seemed to transform a standard apple into a columnar-like tree (Okada et al., 2016). These lines of evidence support that the retroposon induced ectopic expression of the 2OG-Fe(II) oxygenase encoding gene *Co* in shoots and leaves

is biologically responsible for the columnar phenotype.

Despite the advances in revealing the genetic and biological factors underlying the *Co* locus, our current understanding of the genetic control of columnar growth habit remains incomplete. This is because columnar progeny often are observed less than expected in breeding populations segregating for the trait, suggesting there are modifier genes involved (Hemmat et al., 1997; Kenis & Keulemans, 2007; Lapins, 1976; Meulenbroek, Verhaegh, & Janse, 1998). In the present study, we observed that among the 208 F_1 individuals carrying the retroposon insertion in a cross between two columnar selections that are heterozygous at the *Co* locus, 67, 51 and 30 showed standard growth habit respectively in 2-, 4- and 8-year-old trees, indicating there are age-dependent recessive repressors that can suppress columnar phenotype. To identify the columnar repressors, we explored and identified DNA variants of segregation types suitable for mapping recessive traits in apple by pooled genome sequencing, an adaptation from a previous approach developed for mapping dominant traits (Dougherty, Singh, Brown, Dardick, & Xu, 2018). We identified two recessive loci, designated *c2* and *c3* on chromosomes 10 and 9, respectively, which are of significant effect on repressing the columnar phenotype in an age-dependable manner (more significant in young trees than in aged trees). Using RNA-seq analysis, we further revealed that suppressed columnar phenotype is coupled with a drastic expression repression of the *Co* gene, which together with the differentially expressed genes under *c2* and *c3* forms a co-expression gene-network module highly associated with growth habit. Overall, this study represents an important first step towards revealing the identity of the causal genes under *c2* and *c3*, which would greatly increase our understanding of the genetic network responsible for columnar growth habit.

Materials and Methods

Plant materials and growth habit evaluation

The mapping populations were derived from a cross between NY123 (*Coco*) and NY317 (*Coco*) and its reciprocal cross, comprising 246 and 29 (275 in total) F_1 seedling trees, respectively. Since the reciprocal cross' contribution accounted for only 10.5% (29/275), its maternal and paternal effect was ignored. Both parents are advanced breeding selections of columnar growth habit inherited from 'Wijcik McIntosh', the source of columnar apple. The crosses were made in 2007 and the seedling trees were planted in spring 2008 in an experimental orchard of Cornell University, Geneva, New York. The orchard was managed with minimal pruning. Their growth habit was visually evaluated in 2009, 2011, and 2015 based on thickness of stem, number and crotch angle of lateral branches on the main axis and internode length as described previously (T. Bai et al., 2012). Since columnar trees usually have a thicker main stem characterized by similar diameter at the tip and the base, fewer lateral branches with narrower crotch angles, and shorter internodes, the seedling trees were first grouped into two categories: columnar and standard. Next, they were evaluated again based on these characters, dividing each category into two groups, which were scored as "1" for columnar (C) and "2" for columnar-like (CL) in category columnar, and "3" for standard-like (SL) and "4" for standard (S) in category standard for subsequent analysis (**Figure 4.1 A-B**). Columnar-like is a group of columnar that are much taller and/or have a few more branches than a typical columnar, whereas standard-like is a group of standard that have relatively smaller branch angles and/or fewer lateral branches than a typical standard (**Figure 4.1 B**). (Some standard and all standard-like progenies turned out to have a columnar genotype *CoCo* or *Coco*.) Growth habit scores of the eight progenies not determined in 2009 due to small plant sizes were inferred from their scores in 2011, including

five columnar progenies of genotype *Coco* and three standard progenies of *coco* (2) and *CoCo* (1).

Genetic analyses of the *Co* locus

In 2009, the *Co* locus was genotyped with DNA markers SCAR682, EMPc105, 18470-26732, C7629-22009 and HI01a03 in the *Co* region (T. Bai et al., 2012). Unexpectedly, these marker data revealed that 67 of the 208 progenies of genotypes *CoCo* and *Coco* with standard or standard-like growth habit. To differentiate the standard growth habit observed on different *Co* genotypes, the standard growth habit of a non-columnar genotype *coco* is called standard1 (Std1), and that of a columnar genotypes *CoCo* or *Coco* is called standard2 (Std2) (**Figure 4.1 A-B**).

Pooled genome sequencing analysis

Genomic DNA samples from 18 columnar and 16 Std2 progeny (**Figure S4.1**, **Table S4.1**) were isolated in 2015 again and pooled equally with 500 ng each progeny to construct a columnar pool and a Std2 pool, respectively. Sequencing and data analysis of the two genomic pools were conducted similarly as previously described (Dougherty et al., 2018). Briefly, the two pooled genomic DNA libraries of target insert size of 500 bp were constructed using NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA), and then paired-end (2×151 bp) sequenced on an Illumina Nextseq500 platform at the Genomics Facility of Cornell University (Ithaca, New York, USA). The sequencing effort generated 166.5 million and 259.1 million reads in pools Std2 and columnar (NCBI SRA accession SRP200597), covering the reference genome of apple ‘Golden Delicious’ double haploid (comprising 17

chromosomes and a unanchored “chromosome” with a total size of 709.6 Mb) (Daccord et al., 2017) by 35.4× and 55.1×, respectively (**Figure S4.1**, **Table S4.2**). Removing low quality reads and/or bases resulted in 163.5 million 125.0-bp cleaned reads (20.4 Gb) for pool Std2 and 253.0 million 125.8-bp clean reads (31.8 Gb) for pool columnar (**Table S4.2**). Software CLC Genomics Workbench (v11.0, CLCBio, Cambridge, MA, USA) was employed to map the Illumina sequencing reads against the reference genome for each genomic pool. The reads mapping parameters were set as the following: minimum length fraction 0.8 and the minimum similarity 0.98, similar to those described earlier (Y. Bai, Dougherty, & Xu, 2014), leading to mapping 127.5 million (78.0%) of the clean reads in pool Std2 and 195.1 million (77.1%) in columnar, covering the genome by 22.5× and 34.6×, respectively (**Table S4.2**).

DNA variants were called for each pool using the fixed ploidy (2x) variant detection tool of CLC Genomics Workbench, which automatically calculates read coverage and variant frequency. A minimum coverage of ten and a minimum count of two for variant-carrying reads were used initially. Next, the variants were filtered to remove those that are non-single nucleotide variants (SNVs), reference alleles, or with reads coverage lower than 20 or greater than 200, resulting in detection of 1,997,962 SNVs commonly present in both pools, 14,078 SNVs specific to pool Std2, and 56,571 specific to columnar (**Figure S4.1**). Notably, a considerable fraction (14.6-16.2%) of the SNVs common to both pools had a variant allele frequency (AF) ≥85% (close to be homozygous) (**Figure S4.2 A-B**) while 90.6% of the SNVs specific to pool Std2 and 94.0% of the SNVs specific to pool columnar had a variant AF ≤45% (Figure S2C-D). Based on the genetics of recessive traits alongside such distribution of the variant AFs, it was postulated that only the SNVs common to both pools may contain SNVs informative for mapping recessive Std2 (**Figure S4.1**, **Figure S4.2**, **Table S4.3**). This is because

SNVs in progenies carrying a recessive trait (pool Std2) are expected to be mostly homozygous near the causal genes while a majority of the SNVs in progeny without the trait (pool Columnar) would be heterozygous. The SNVs specific to either pools are non-informative for mapping recessive Std2 although those specific to pool columnar could be informative for mapping the alleles dominant over Std2 (**Figure S4.1**, **Figure S4.2**, **Table S4.3**).

Inferring informative variant segregation types for mapping recessive trait

The strategy for inferring variant segregation types informative for mapping a dominant trait described before (Dougherty et al., 2018) was followed. However, it was adapted to recessive Std2, which is necessary due to recessive inheritance (see also Discussion). The task here is to identify variants that not only situate with high density in the genomic regions targeted by phenotypic pooling, but also differ widely between pools Std2 and columnar in variant allele frequency due to their physical linkage to the causal genes and due to their segregation types. To facilitate the identification of informative variants for mapping, the SNVs (1,997,962) common to both pools were grouped into heterozygotes (15%≤AF<85%) and homozygotes (AF≥85%). Combining their allele frequency with source pools, the SNVs were further divided into four groups (**Table S4.3**): 1) 70,522 (3.5%) homozygotes in Std2 and heterozygotes in columnar (Ho-Std2/He-Col); 2) 1,636,085 (81.9%) heterozygotes in both Std2 and columnar (He-Std2/He-Col); 3) 39,075 (2.0%) heterozygotes in Std2 and homozygotes in columnar (He-Std2/Ho-Col); and 4) 252,280 (12.6%) homozygotes in Std2 and homozygotes in columnar (Ho-Std2/Ho-Col).

In a typical bi-parental cross in apple, the segregation of SNVs could be determined by up to six possible segregation types. They include <ab x cd>, <ef x eg>, <hk x hk>, <lm x ll>, and <nn x np> and <qq x qq>, where each letter stands for one of the four DNA bases (A, C, G and T) in SNVs (Dougherty et al., 2018). Among them, variants of the segregation type <qq x

qq> clearly are non-informative due to non-segregating, whereas those in the other five types are considered informative for mapping Std2. However, variants of segregation types <ab x cd> and <ef x eg> are rare in the genome due to their involvement of four or three DNA bases, thereby not further pursued. The remaining variants of segregation types <hk x hk>, <lm x ll> and <nn x np>, which involve only two nucleotides, are more abundant and suitable for mapping. When the SNVs' linkage to the Std2 allele (i.e. haplotype) is considered, the three suitable segregation types could be expressed with at least 12 derivatives (**Table S4.3**).

Examining the allele frequency of SNVs under each of the 12 possible segregation types under the model of one- or two-recessive genes revealed five variant segregation types potentially informative and suitable for mapping recessive Std2. They include <**h**k × **h**k>, <**l**m × **ll**>, <**nn** × **n**p>, <**l**m × mm>, and <pp × **n**p>, named segregation types A, B, C, D and E, respectively. Here the letters in bold present SNVs in relation to the reference genome, and the alleles in each first and third positions are assumed in linkage with the recessive Std2 alleles in the seed and pollen parents, respectively (**Figure S4.3**, **Table S4.3**). The others seven segregation types were not informative for mapping recessive Std2 due to either an equal SNV allele frequency in both pools or a negative SNV allele frequency margin (informative for mapping the allele dominant over Std2) between pools Std2 and columnar.

Variants under segregation types A, B and C were inferred to be homozygous in pool Std2 while heterozygous in columnar (Ho-Std2/He-Col). Considering the cases involving one- or two- recessive genes, the variant allele frequency in pool columnar would be 33.3% and 46.7% for type A (<**h**k × **h**k>), and 66.7% and 73.3% for both types B (<**l**m × **ll**>) and C (<**nn** × **n**p>), respectively. The variant allele frequency directional (positive) difference (AFDD) between pools Std2 and columnar would be 66.7 and 53.3 percentage points for type A, and 33.3 and 26.7

percentage points for both types B and C, respectively (**Figure S4.3**, **Table S4.3**).

Similarly, variants under segregation types D (<**l**m × mm>) and E (<pp × **n**p>) were inferred to be heterozygous in both pools, but different in their variant allele frequency. Under the model of one- or two- recessive genes, the allele frequency in pools Std2 and columnar would be 50.0% and 16.7%, and 50.0% and 23.3%, corresponding to AFDD 33.3 and 26.7 percentage points, respectively (**Figure S4.3**, **Table S4.3**).

These analyses confirmed that only the variants common to both pools could be informative for mapping the recessive Std2 trait while none of the pool specific variants would be informative. Within the pool common variants, only those that are in groups Ho-Std2/He-Col and He-Std2/He-Col are useful potentially (**Figure S4.1**, **Table S4.3**).

Identification of informative SNVs for mapping recessive Std2

The allele frequency directional (positive) difference (AFDD) and density (AFDDD) AFDDD mapping approach, which explores DNA variants that are common to both pools as previously described (Dougherty et al., 2018), was adapted to map recessive Std2. This was accomplished by identifying informative SNVs based on the expected variant allele frequencies in each pool and their expected directional (positive) differential margins between pools Std2 and columnar under each of the five informative segregation types (**Figure S4.1**, **Figure S4.3**; **Table S4.3**, **Table S4.4**). For segregation type A (<**h**k × **h**k>), the 70,522 SNVs in group Ho-Std2/He-Col were subjected to two filters: 1) Variant allele frequency ≥85% in pool Std2. 2) The AFDD ≥43.3 percentage points between pools Std2 and columnar to cover the expected AFDD 66.7 and 53.3 (percentage points) under the model of one- and two-recessive genes, respectively (Table S4). For segregation types B (<**l**m × **ll**>) and C (<**nn** × **n**p>), the 70,522 SNVs were

filtered similarly as described above. However, the AFDD was limited to a range from 16.7 to 43.3 percentage points between pools Std2 and columnar (Table S4). For segregation types D (<**l**m × mm>) and E (<pp × **n**p>), the 1,636,085 variants that are in group He-Std2/He-Col were filtered with two filters: 1) the DNA variant AF ranges from 35% to 65% in the Std2 pool, close to their estimated mean 50%. 2) The AFDD is restricted from 16.7 to 43.3 percentage points between the Std2 and columnar pools (**Table S4.4**).

As a result, 7,642 informative variants were identified under segregation type A, 40,166 under types B and C, and 70,230 under types D and E. These informative variants (SNVs), either combined or individually according to their segregation types, were then plotted along the reference genome and visualized using a total number of variants in 1-Mb sliding windows. Genomic regions of variant density significantly higher the genome average in standard score (z) test were consider putatively linked to the recessive trait Std2. The z-test was conducted in MS-Excel or R if the p values were lower than 1.0E-7, and the cutoff p value (two-tailed confidence level) is $-\log_{10}p(z)$ (called LODz for convenience) >2.5.

Development and analysis of DNA markers in genomic regions putatively linked to Std2

As a validation step, independent DNA markers were used to confirm the putative genetic linkage between trait Std2 and the *c2, c3,* and other positive regions identified by AFDDD mapping in pooled genome sequencing analysis. Existing SSR markers in these regions were applied first. If necessary, new SSRs would be developed from the apple reference genome as described earlier (Xu, Wang, & Brown, 2012). Polyacrylamide gel electrophoresis of SSR markers were conducted as detailed previously (Wang et al., 2012). In addition, high-resolution melting (HRM) markers were developed by targeting SNVs of segregation type A-C in coding

regions of genes. Analysis of HRM markers was performed using a CFX96 Touch Real-Time PCR Detection System in combination with Precision Melt Super Mix and software packages CFX Maestro and High Resolution Melting Analysis following the manufacturer's instruction (BioRad, Hercules, California). The SSR and HRM marker primer sequences and their approximate physical locations in the reference genome were provided (**Table S5.5**).

RNA-seq analysis

Twelve F_1 progenies evenly representing phenotypes columnar (4), Std1 (4) and Std2 (4) were budded onto apple rootstock B118 in 2015 and planted in the orchard in spring 2016 (Figure 1A-B). In June 2017, the actively growing shoot apex tissues (leaves removed) were collected and flash frozen in liquid nitrogen for RNA isolation. The shoot apex tissues were ground in liquid nitrogen and total RNA was isolated as previously described (Meisel et al., 2005). The RNA samples were treated with DNase I (amplification grade, Invitrogen, Carlsbad, CA) and cleaned with RNeasy MinElute Clean up Kit (Qiagen, Hilden, Germany). RNA concentration and quality were determined using NanoDrop 1000 (Thermo Fisher Scientific, Waltham, MA) and assays on a 1.0% agarose gel.

mRNA was isolated from total RNA using NEBNext Poly(A) mRNA Magnetic Isolation Module and was used to construct RNA-seq libraries with NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) as previously described with slight modification (Y. Bai et al., 2014). Briefly, libraries were size-selected for 350-500bp and were multiplexed in equal amount for single end 76-base sequencing by NextSeq 500 (Illumina, San Diego, CA) at the Cornell University Biotechnology Resource Center (Ithaca, New York). Illumina sequencing of the pooled RNA-seq libraries generated 12 FASTQ files of sequences

(NCBI SRA accession SRP200597) with 409.9 million raw reads in total (**Table S4.6**). Cleaning-up of the raw reads, including removal of adaptors, rRNA contaminations and low quality and/or short reads, was conducted similarly as described previously (Y. Bai, Dougherty, Cheng, Zhong, & Xu, 2015). The resultant high quality reads were mapped to the apple reference genome (Daccord et al., 2017) using CLC Genomics Workbench 11.0 (Minimum similarity fraction: 0.98, minimum length fraction: 0.8, and maximum number of hits: 10). Gene expression levels were calculated and represented by reads per kilobase of exon model per million mapped reads (RPKM) (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008). Genes were considered expressed when their mean RPKM>0.25 in any of the three sample groups (columnar, Std1 and Std2). Differentially expressed genes (DEGs) were defined as those of RPKM folder change≥1.5 and P_{FDR}≤0.05 among the three groups.

Weighted gene co-expression network analysis (WGCNA)

DEGs among phenotype groups columnar, Std1 and Std2 were analyzed using WGCNA, an R package (Langfelder & Horvath, 2008) to identify co-expression gene network modules associated with growth habit. The significance cutoff is p0.001. Relevant parameters were set similarly as described previously (Y. Bai et al., 2015). Visualization of the most significant WGCNA module was accomplished using Cytoscape 3.6 (Saito et al., 2012). Analyzing the network was carried out using a Cytoscape plugin Network Analyzer (Assenov, Ramírez, Schelhorn, Lengauer, & Albrecht, 2008).

MapMan annotation and gene enrichment analysis

Annotations of the reference genome with MapMan Bins was assisted with Mercator

(Lohse et al., 2014), resulting in assigning a MapMan bin to 45,116 genes. Gene enrichment analysis was performed for the WGCNA module that shows the highest correlation with tree growth habit using the hypergeometric annotation test tool available in CLC Genomics Workbench, which is similar to the unconditional GOstats test (Falcon & Gentleman, 2007). For declaration of significant enrichment, the cutoff is $P_{FDR}<0.05$.

Quantitative (q) RT-PCR

The same set of plant samples taken in June 2017 for RNA-seq were used, and another set taken in June 2018 was used to repeat the analysis for the *Co* gene. Two microgram of total RNA was used in reverse transcription reactions using the iScript™ cDNA Synthesis Kit (BioRad, Hercules, CA) to obtain the first strain of cDNA, and then used as templates for qRT-PCR analysis. The qRT-PCR reactions were performed with three technical replicates using iTaq™ Universal SYBR® Green Supermix on a CFX96 Touch Real-Time PCR Detection System according to manufacturer's protocol. An apple actin encoding gene (MD01G1001600) was used as a reference gene. The expression levels of target genes were quantified based on the normalized expression (ΔΔCq) of the reference gene actin using the Bio-Rad CFX Maestro software. qRT-PCR primers were designed for eight genes expressed in the libraries (Table S5). The normalized gene expression from qRT-PCR was compared to the RNA seq gene expression in RPKM.

Statistical analysis

Analysis of variance (ANOVA) and regression analysis were conducted using JMP Pro12 (SAS, Cary, NC).

Results

Segregation of columnar and standard phenotypes

The 275 F_1 progenies derived from (reciprocal) cross NY123 (*Coco*) × NY317 (*Coco*) were planted in 2008 and were genotyped with a codominant DNA marker 18470-25831 co-segregating with columnar (T. Bai et al., 2012) and five other markers SCAR682, EMPc105, C7629-22009 and 29f1/JWI1r that link to columnar or detect the retroposon insertion (T. Bai et al., 2012; Wolters et al., 2013). The marker genotypic data confirmed a normal 1:2:1 segregation (χ^2=1.029, p=0.5978) for genotypes *coco (*67), *Coco* (132), and *CoCo* (76) in the F_1 populations. However, evaluation of their growth habits in 2009, 2011, and 2015 indicated that the expected 3:1 (columnar: standard) segregation was significantly distorted (χ^2=82.46, p=2.2E-16 in 2009; χ^2=47.04, p=7.0E-12 in 2011; χ^2=16.49, p=4.9E-5 in 2015) for columnar (C) and columnar-like (CL) vs. standard (S) and standard-like (SL) due to excessive S/SL individuals. Inspecting the segregation data (Figure 1C-E) indicated the following: 1) All 67 seedlings of genotype *coco* were consistently standard. 2) In the 208 progenies of genotypes *CoCo* and *Coco*, 141 consistently exhibited C/CL as expected during the period studied. 3) The remaining 67 individuals of genotypes *CoCo* and *Coco* were scored as standard and standard-like unexpectedly in 2009, and were progressively reduced to 51 in 2011, and 30 in 2015, i.e. 37 of the 67 S/SL progenies in 2009 progressively returned to C/CL while the other 30 remained unchanged (**Figure S4.1**).

These observations suggested that the presence of the 67 to 30 S/SL individuals of genotypes *CoCo* and *Coco* directly caused the phenotypic segregation distortion. For convenience, the standard phenotype associated with genotype *coco* is called standard1 (Std1), and that with *CoCo* and *Coco* called standard2 (Std2) (**Figure 4.1 A-B**). Since the Std2

individuals accounted for 14.4 to 32.2 percent in the *CoCo* and *Coco* genotype groups, there are age-dependent recessive repressors (genes) that can repress the columnar phenotype more effectively in young trees than in aged trees in the populations.

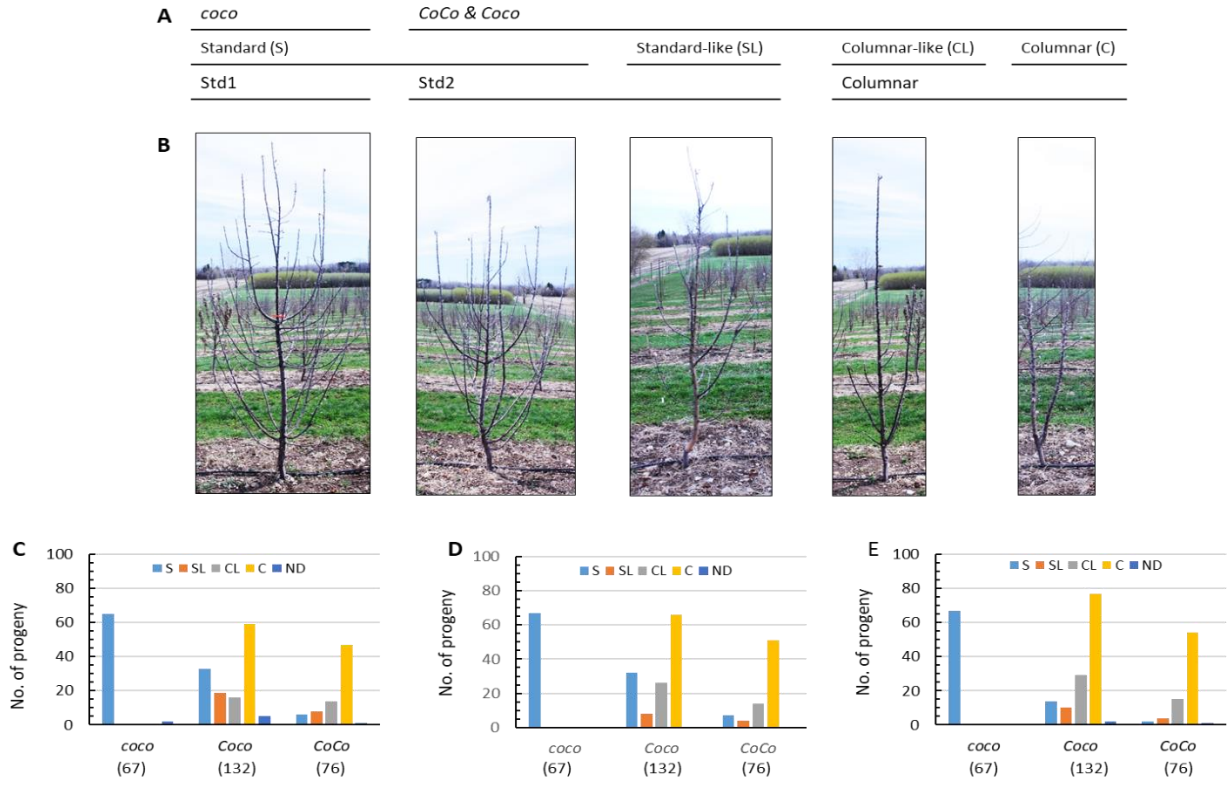


Figure 4.1. Growth habit evaluation. (**A**) Diagram of the genotypic and phenotypic relationships among Std1, Std2 and columnar. (**B**) Representative trees of standard (S), standard-like (SL), columnar (C) and columnar –like (CL) growth habits. The pictures were taken from three-year-old budded trees on rootstock B118. (**C**-**E**) Observed genotypes at the Co locus in C, CL, S and SL in 2009 (**C**), 2011 (**D**) and 2015 (**E**). ND: not determined.

Pooled genome sequencing based AFDDD mapping of the recessive repressors of columnar

To genetically map the columnar recessive repressors without construction of the linkage map in the population, the pooled genome sequencing based variant allele frequency directional (positive) difference and density (AFDDD) mapping approach (Dougherty et al., 2018) was adapted for recessive traits. The first pool comprised 16 progenies of Std2 growth habit; whereas

the second was formed with 18 progenies of columnar growth habit (**Table S4.1**). Subsequent pooled genome sequencing and detailed DNA variant analyses (**Figure S4.1, Figure S4.2, Figure S4.3; Table S4.2, Table S4.3, Table S4.4**) identified 7,642 SNVs under informative segregation type <**h**k × **h**k> (type A), 40,166 SNVs under <**l**m × **ll**> (types B) and <**nn** × **n**p> (type C), and 70,230 SNVs under <**l**m × mm> (types D) and <pp × **n**p> (type E). Here each letter denotes one of the four DNA bases and the letters in bold present SNVs in relation to the reference genome. The alleles in each first and third positions are assumed in linkage with the recessive Std2 alleles in the seed and pollen parents, respectively (**Figure S4.3**, **Table S4.3**).

Examining the genome distribution of the three sets of informative SNVs collectively (118,038 in total) revealed five genomic regions of significantly higher variant density that the genome average, named *c2, c3, c4, c5 and c6*, respectively (**Figure 4.2 A**). The peaks of *c2* and *c3* are located at the 27th (26-27) Mb on chromosome 10 (LODz=29.2), and the 15th Mb on chromosome 9 (LODz=12.2), respectively (**Figure 4.3**). The peak locations of *c4*, *c5* and *c6* were at 26th Mb on chromosome 14 (LODz=3.92), the 4th Mb on chromosome 6 (LODz=3.04), and the 27th Mb on chromosome 8 (LODz=2.51), respectively (**Figure 4.3**). Clearly, the *c2* and *c3* regions likely represent the major loci relevant for phenotype Std2.

To see if the five variant segregation types may contribute differently to the identified regions, the genome distribution of the three sets of variants were examined independently (**Figure 4.2 B-C**). The results demonstrated that two major loci *c2* and *c3* were determined by SNVs of segregation types A, B and C, *c5* and *c6* by those of types D and E, and *c4* by all segregation types. Therefore, segregation types A, B and C appeared to be more useful than types D and E in this study.

Figure 4.2. Variants allele frequency (AF) directional difference (AFDD) and density (AFDDD) mapping of columnar recessive repressors using the five informative segregation types of 118,038 SNVs (**A**), type A <hk × hk> of 7,642 SNVs (**B**), types B-C <lm × ll> and <nn × np> of 40,166 SNVs (**C**), and types D-E <pp × np> and <lm × mm> of 70,230 SNVs (**D**). The numbers on X-axis represents apple chromosomes. The line in red dashes indicates the cutoff LODz (-log10p(z)) 2.5 in z-score test of AFDDD in (**A**).

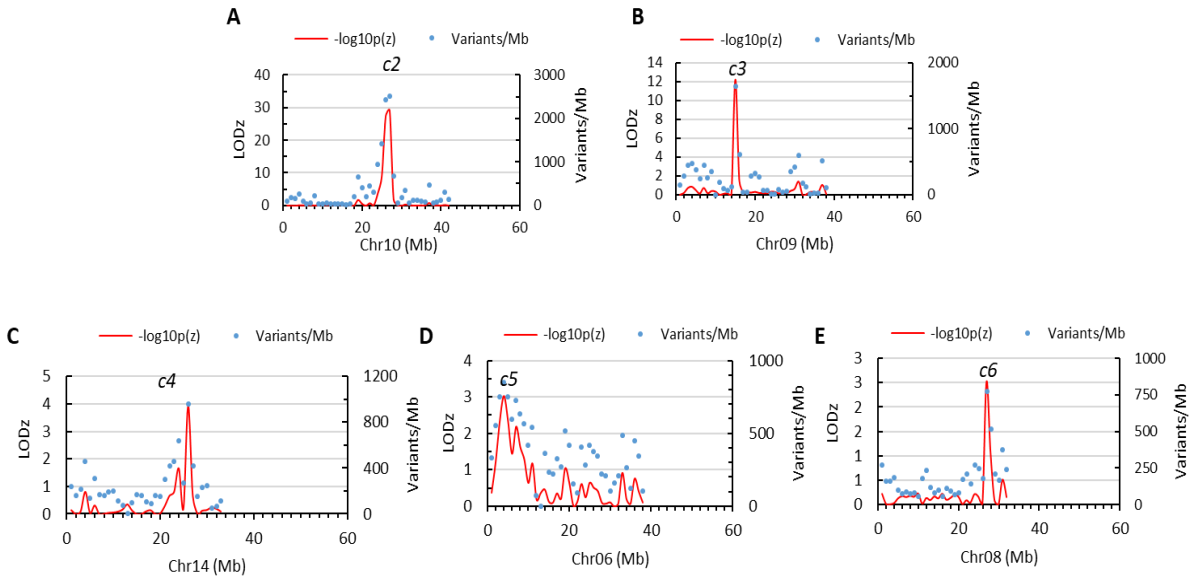


Figure 4.3. Close-up views of the genomic regions mapped by AFDDD mapping for putative columnar recessive repressors *c2* (**A**), *c3* (**B**), *c4* (**C**), *c5* (**D**) and *c6* (**E**). Note that the *c2* region is close to the Co gene, which located at 28.0th Mb on chromosome 10.

Confirmation of the mapping of Co repressors

Identification of loci *c2* to *c6* indicates putative mapping of the recessive repressors. For confirmation, 13 existing and newly developed SSR and HRM markers in these regions were analyzed in the populations (**Table S4.5**). To maximize the confirmation, the 2009 phenotypic data were used. Based on the marker-trait linkage analysis in the 208 progenies of genotypes *CoCo* and *Coco*, loci *c4 to c6* were not confirmed (data not shown) while loci *c2* and *c3* were confirmed, which were described below.

The *c2* locus was represented by marker AU223548, which is located at the 26.35th Mb on chromosome 10, roughly 1.65 Mb upstream of the *Co* gene (MD10G1185400) encoding a 2OG-Fe(II) oxygenase (Okada et al., 2016; Otto et al., 2014; Wolters et al., 2013). The observed recombination rates between the two loci were 0.079 in NY123 and 0.113 in NY317, suggesting a moderate linkage between them. Since the repression effect of *Co* could not be detected in the

coco group, the genetic effect of *c2* was investigated only in genotype groups *CoCo* and *Coco*. The null hypothesis is that the *c2* and *C2* alleles from a given parent may segregate differently between genotype groups *coco* and *CoCo/Coco* due to linkage, but would segregate similarly across *CoCo* and *Coco* irrespective of their phenotypes columnar and Std2. A significant segregation distortion from what is expected in phenotype group Std2 and/or columnar would indicate a linkage between *c2* and Std2. The parental recessive *c2* alleles were defined as those whose frequencies were increased significantly in Std2.

Chi-square analysis of marker AU223548 genotypes indicated that the segregation of parental alleles was significantly distorted in Std2 (p=0.0011 in NY123; p=6.40E-6 in NY317) (**Figure S4.4 A, B**) and columnar (p=0.0256 in NY123; p=0.0021 in NY317) (**Figure S4.4 D, E**). Consequently, the *c2c2* progenies were significantly more frequent (p=6.28E-10) than what was expected in Std2 while the *C2C2* individuals were significantly more (p=1.17E-6) in columnar (**Figure S4.4 C, F**). These observations supported strongly the genetic mapping of locus *c2*. Examining the relationships between loci *Co* and *c2* revealed that NY123 and NY317 are of genotypes *Coc2* *coC2* and *CoC2* *coc2* (the underlines denote haplotype, and the genotypes are signified with alleles from NY123 and NY317 in order), respectively. In other words, the recessive *c2* allele is linked to the dominant *Co* allele in coupling phase in NY123 while in repulsion phase in NY317.

The *c3* locus was confirmed by marker Hi05e07, which is located at the 14.3th Mb on chromosome 9. The marker segregation was distorted for NY317 alleles in Std2 (p=0.0055) and columnar (p=0.0546) (**Figure S4.4 H, K**), but was normal for the NY123 alleles (p=0.1894 in Std2 and 0.3641 in columnar) (**Figure S4.4 G, J**). This suggested that NY317 and NY123 are heterozygous and homozygous at the *c3* locus, respectively. The genotype of NY123, therefore,

is inferred as *Coc2 coC2 c3c3*, and that of NY317 as *CoC2 coc2 C3c3*. The distorted segregation of NY317 alleles led to a significant increase in the number of *c3c3* individuals in Std2 (p=0.0189) although the increase for the *C3C3* progenies were not significant (p=0.1232) in columnar (**Figure S4.4 I, L**).

Investigating the phenotypic frequencies in each of the six possible genotype groups (three genotypes *Coc2 CoC2, Coc2 coc2,* and *coC2 CoC2* at *c2* by two genotypes *c3c3* and *c3C3* at *c3*) further supports the association between phenotype Std2 and recessive genotypes *c2c2* and *c3c3* (Figure 4). The six genotypes are *c2C2 c3C3, c2C2 c3c3, c2c2 c3C3, c2c2 c3c3, C2C2 c3C3* and *C2C2 c3c3* when omitting the *Co* alleles. In 2009, the frequencies of Std2 were high in double recessive carriers *c2c2 c3c3* (0.867), medium in single recessive *c2c2 c3C3* (0.478), *c2C2 c3c3* (0.308) *and C2C2 c3c3* (0.182), and low in non-recessive *c2C2 c3C3* (0.156) and *C2C2 c3C3* (0.042). The overall frequency of Std2 was reduced by tree age; however, the trend remained (**Figure 4.4**). By 2015, the frequencies of Std2 in double recessive, single recessive and non-recessive were 0.448, 0.136-0.154 and 0-0.032, respectively. These observations also suggested that the penetrance of phenotype Std2 was incomplete even in double recessive *c2c2 c3c3*, ranging from 0.867 in 2009 to 0.448 in 2015.

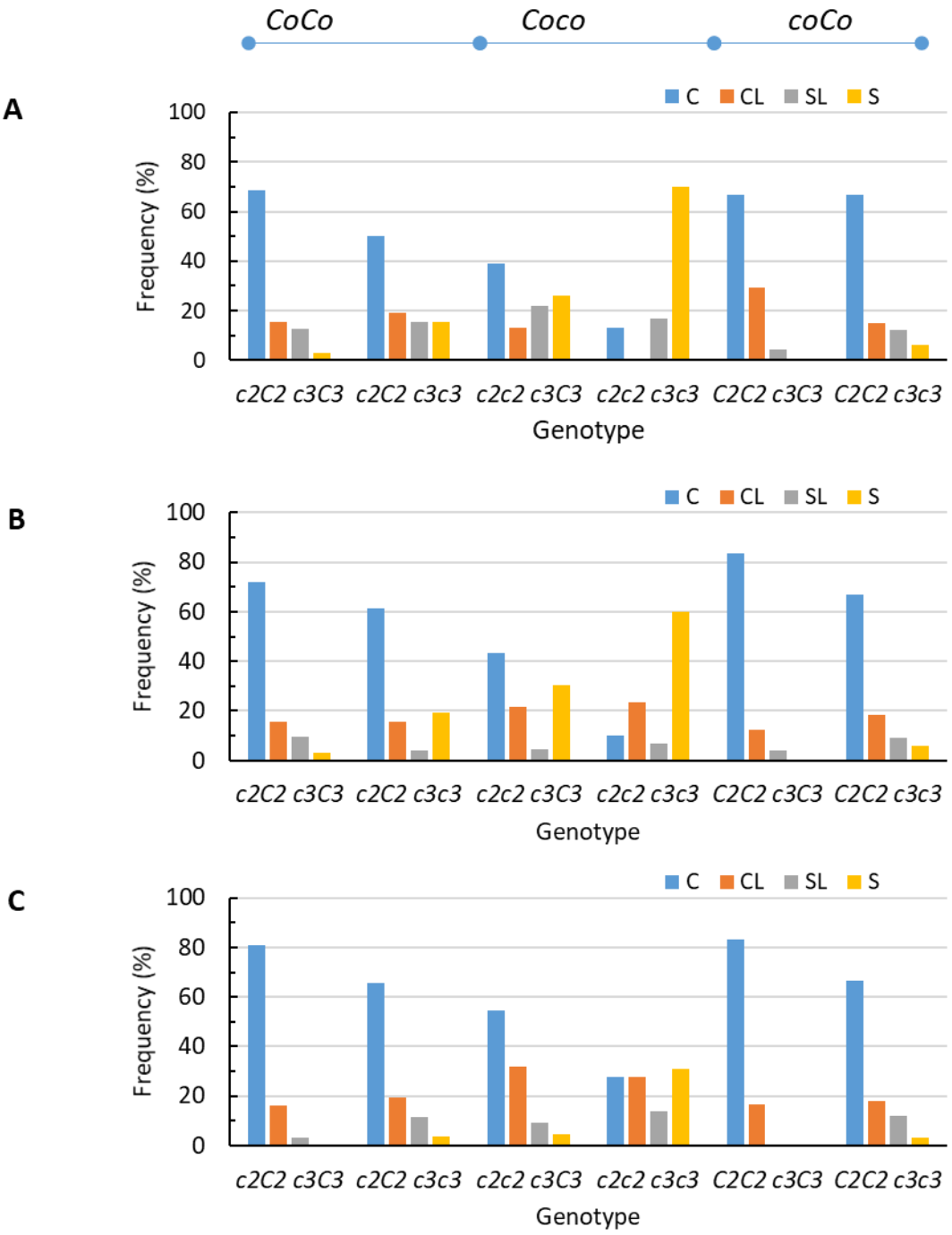


Figure 4.4. Phenotypic frequencies in each of the six possible genotypes observed in years 2009 (**A**), 2011 (**B**) and 2015 (**C**) when recombinants between *Co* and *c2* were excluded due to limited number for meaningful comparisons. The six genotypes are *Coc2 CoC2 c3C3*, *Coc2 CoC2 c3c3*, *Coc2 coc2 c3C3*, *Coc2 coc2 c3c3*, *coC2 CoC2 c3C3*, and *coC2 CoC2 c3c3*, i.e. *c2C2 c3C3*, *c2C2 c3c3*, *c2c2 c3C3*, *c2c2 c3c3*, *C2C2 c3C3* and *C2C2 c3c3* if omitting *Co* alleles. The underlines denote haplotypes, and the genotypes are signified with alleles from NY123 and NY317 in order. C: columnar; CL: columnar-like; S: standard; SL: standard like.

Genetic effect of c2 and c3 on repression of columnar

To quantify the genetic effect of *c2* (AU223548) and *c3* (Hi05e07) on repression of the columnar phenotype in genotype groups *CoCo* and *Coco*, regression analyses were conducted by assigning the phenotypes columnar, columnar like, standard like and standard with scores 1, 2, 3 and 4, respectively. The results revealed that locus *c2* accounted for 19.2% of the phenotypic variation in 2009, 16.2% in 2011, and 10.0% in 2015, greater than what was estimated for locus *c3*, which are 8.1%, 6.9% and 7.0%, respectively, and the two loci combined explained 25.7%, 22.2% and 15.7% of the population variance in 2009, 2011 and 2015, respectively (**Table 4.1**). The regression model fit p-values ranged from 4.93E-12 to 6.68E-04, which were all significant (**Table 4.1**). Overall, *c2* seemed to play a much greater role in repression of *Co* in younger trees than in older trees while the influence of *c3* was constant.

Table 4.1. Regression analyses of the effect of loci *c2* and *c3* on repression of columnar phenotype

loci	2009			2011			2015		
	r^2	p	n	r^2	p	n	r^2	p	n
c2 (AU223548)	0.1921	6.02E-10	202	0.1619	1.37E-08	208	0.0997	2.48E-05	205
c3 (Hi05e07)	0.0806	2.34E-04	202	0.0692	6.68E-04	207	0.0702	6.66E-04	204
c2 and c3	0.2574	4.93E-12	202	0.2219	2.32E-10	207	0.1571	6.88E-07	204

P (for model fit) is determined by ANOVA.

Transcriptomic characterization of main shoot apex in columnar, standard1 and standard2

An RNA-seq analysis was conducted on actively growing main shoot apex tissues from four columnar, four Std1 and four Std2 progenies grafted on rootstock B118 (**Figure 4.5 A**, **Table S4.6**) to investigate what genes were expressed differentially genome-wide and locally in the *c2* and *c3* regions and how the *Co* gene (MD10G1185400) was repressed in the three groups. After removal of low quality reads and rRNA contaminations, 334.2 million clean reads (76 bp) in total were obtained from the 12 libraries, and 279.3 million (83.6%) of them were mapped to the apple reference genome (Daccord et al., 2017), equivalent to 23.3±9.7 (83.7±1.2%) million mapped reads per sample (**Table S4.6**). In total, there were 33,430 genes expressed (mean RPKM≥0.25 in columnar, Std1 or Std2). Principle component (PC) analysis of the gene expressions revealed that samples in Std1 form a tight group while samples in Std2 and columnar appeared to form their own groups as well despite more spreading (**Figure 4.5 B**). Pair-wised comparison among the three groups identified 588 differentially expressed genes (DEGs) between Std2 and columnar, 2142 between Std2 and Std1, and 3280 between columnar and Std1 (**Figure 4.5 C**, **Table S4.7**), suggesting Std2 resembles columnar more than Std1, consistent with the results in PC analysis (**Figure 4.5 B**). Venn diagram analysis showed that there were 4,143 non-redundant DEGs among the three groups (**Figure 4.5 C**).

In the 588 DEGs between Std2 and columnar, 392 (66.7%) were down regulated in Std2 while 196 (33.3%) upregulated. In contrast, 1667 (50.8%) of the 3280 DEGs between Std1 and columnar were down regulated in Std1 while 1613 (49.2%) upregulated (**Figure S4.5**), suggesting that a higher proportion of the DEGs were downregulated in Std2 than in Std1.

Validation of RNA-seq based expression was conducted by qRT-PCR analysis on eight genes (**Figure S4.6 A-H**). Highly significant correlations in gene expression were observed between qRT-PCR and RNA-seq (R^2=0.5217 to 0.9479; p=7.98E-3 to 9.66E-8; n=12), indicating the RNA-seq data are reliable.

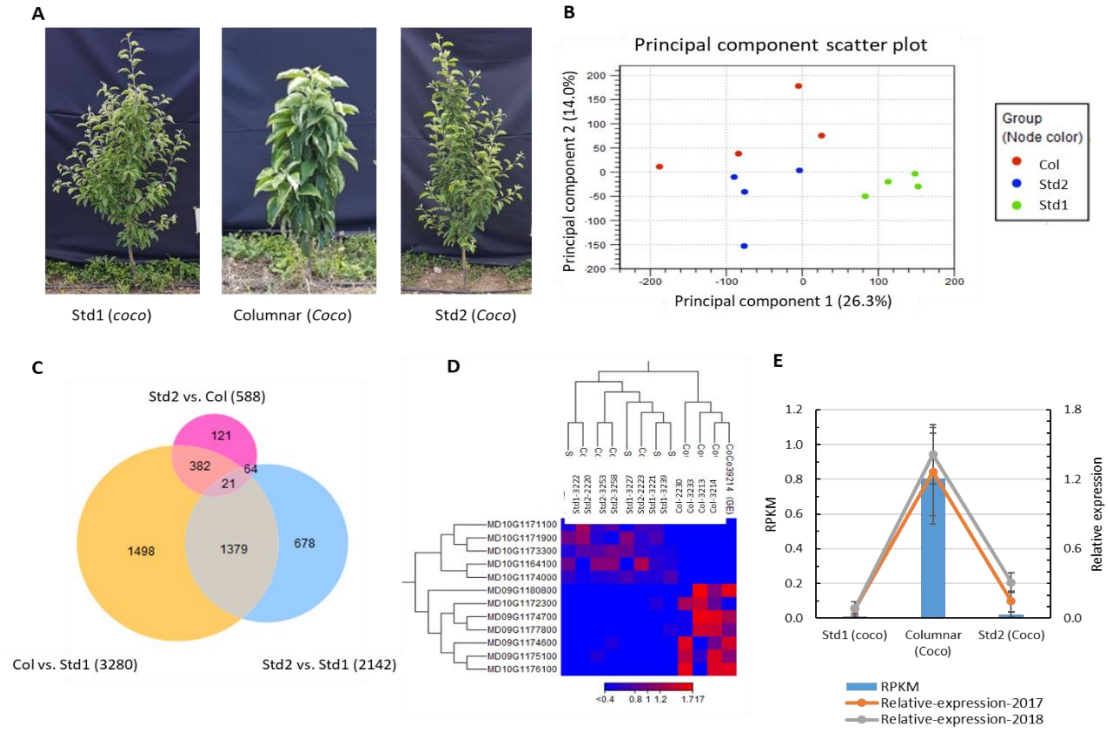


Figure 4.5. Differentially expressed genes (DEGs) in actively growing main shoot apex tissues among the columnar, Std1 and Std2 progenies. (**A**) Photos of typical trees sampled for RNA-seq analysis. The trees were two-year-old (in 2017) budded on rootstock B118, and their main shoot apex tissues were taken for RNA isolation. (**B**) Principal component analyses of 12 RNA-seq samples of columnar (Col), standard1 (Std1) and Standard2 (Std2) growth habits. (**C**) Venn Diagram analysis of the DEGs among the three phenotypes. The numbers in parenthesis indicate the sum of DEGs in each comparison. (**D**) DEGs in the genomic regions of c2 and c3. (**E**) Expression repression of the Co gene in Std2. Relative expression levels were determined by qRT-PCR from samples (n=3x4) taken in 2017 and 2018, respectively. RPKM: reads per kilobase of transcript per Million mapped reads.

Differentially expressed genes under *c2* and *c3*, and repression of *Co* in Standard2

In the *c2* (25.0-27.0 Mb) and *c3* (14.0-16.0 Mb) peak regions (**Figure 4.3 A-B**), there were 155 and 173 genes annotated, of which 109 and 101 were expressed, respectively (**Table S4.8**). Despite the large number of annotated and expressed genes, the DEGs between columnar and Std2 were limited to seven under *c2* and five under *c3* (**Figure 4.5 D, Table S4.8**). Of the seven DEGs under *c2*, two genes MD10G1172300 (encoding a glutathione S-transferase TAU 8-like) and MD10G1176100 (Long-chain fatty alcohol dehydrogenase family protein) were

downregulated in Std2, whereas the other five were upregulated, including MD10G1171100 encoding a GDSL lipase, and MD10G1171900, MD10G1173300 and MD10G1174000 of unknown function. The five DEGs under *c3* were all downregulated in Std2, including MD09G1174600, MD09G1174700 and MD09G1175100 encoding a GDSL lipase, MD09G1177800-an aldolase-type TIM barrel family protein, and MD09G1180800-a protein of unknown function (**Figure 4.5 D**, **Table S4.8**). These DEGs are considered important candidate genes as the columnar repressors. Interestingly, four of the 12 DEGs are GDSL-like genes.

The expression of the *Co* gene (MD10G1185400) was relatively low in columnar (RPKM 0.805±0.262), but clearly detectable. Surprisingly, its expression in Std2 was reduced by 27.8 fold to RPKM 0.021±0.014, close to RPKM 0.008±0.017 in Std1 that was virtually undetectable, suggesting a drastic repression of *Co* (**Figure 4.5 E**, **Figure S4.6 A**). These expression patterns were also detected in qRT-PCR analyses using the same or similar shoot apex tissues collected in 2017 and 2018 (**Figure 4.5 E**). Since the induced higher expression of *Co* by the retroposon insertion is responsible for the columnar phenotype (Okada et al., 2016; Otto et al., 2014; Wolters et al., 2013), the repression of *Co* expression may have suppressed columnar, leading to the Std2 phenotype.

To search for SNVs that could potentially lead to recessive Std2, the 155 genes annotated under *c2* and 173 under *c3* were also investigated for the presence of non-synonymous mutations that are among the 47,808 SNVs under segregation types A-C (**Table S4.4**), which are homozygous in pool Std2 (**Table S4.3**). It is revealed that 58 and 25 expressed genes under *c2* and *c3* carry at least one non-synonymous SNV, respectively (**Table S4.8**). Interestingly, such mutation-carrier genes include two (MD10G1164100 and MD10G1176100) of the seven DEGs under *c2* and three (MD09G1174600, MD09G1174700 and MD09G1175100) of the five DEGs

under *c3* (**Table S4.8**).

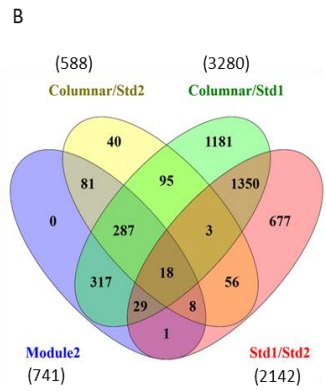
Identification of a *Co* guided co-expression gene network module

Weighted gene expression network analysis (WGCNA) of the 4143 DEGs identified ten WGCNA network modules. Among them, module2 of 741 member genes showed the highest module-trait (growth habit) association (r=0.86, p=0.0003) in the 12 samples (**Figure 4.6A**, **Table S4.7**). A majority (92.4% or 685) of the 741 member genes comprised DEGs from three groups, including 81 (10.9%) DEGs unique to the comparison between columnar and Std2 (columnar/Std2), 317 (42.7%) DEGs unique to columnar/Std1, and 287 (38.7%) DEGs common to both columnar/Std2 and columnar/Std1 (**Figure 4.6 B**, **Table S4.7**). The remaining 56 (7.6%) were from four groups related Std1. On average, the member genes have 431.9±180.7 edges, ranging from one to 708, in module2. Importantly, the *Co* gene (MD10G1185400) is a member of module2, which is connected by 468 primary neighbor genes (**Figure 4.6 C**, **Table S4.7**) that also include five of the seven DEGs under *c2* and four of the five DEGs under *c3* (**Figure 4.5 D**, **Table S4.8**), supporting that module2 represents an important gene network responsible for growth habit.

Compared with columnar samples, 639 (86.2%) of the 741 member genes in module2 were downregulated in Std2 while 73 (9.9%) were upregulated and 29 (3.9%) were unchanged (absolute fold change <1.50). Similarly, 581 (78.4%) of the 741 DEGs were downregulated in Std1, 90 (12.1%) upregulated and 70 (9.4%) unchanged (**Table S4.7**). These observations indicated that module2 member genes were mostly downregulated in Std1 and Std2 (**Figure S4.7**), a trend similarly noted in the 588 DEGs between columnar and Std2 (**Figure S4.5**).

A

Module	Correlation with growth habit (R)	p	No. of member DEGs
module1	-0.44	0.20	2425
module2	**0.86**	**3.0E-04**	**741**
module3	0.66	0.02	466
module4	-0.71	0.01	164
module5	-0.29	0.40	76
module6	-0.55	0.06	76
module7	0.65	0.02	57
module8	-0.58	0.05	55
module9	0.41	0.20	41
module10	-0.39	0.20	39
module0	-0.45	0.10	3



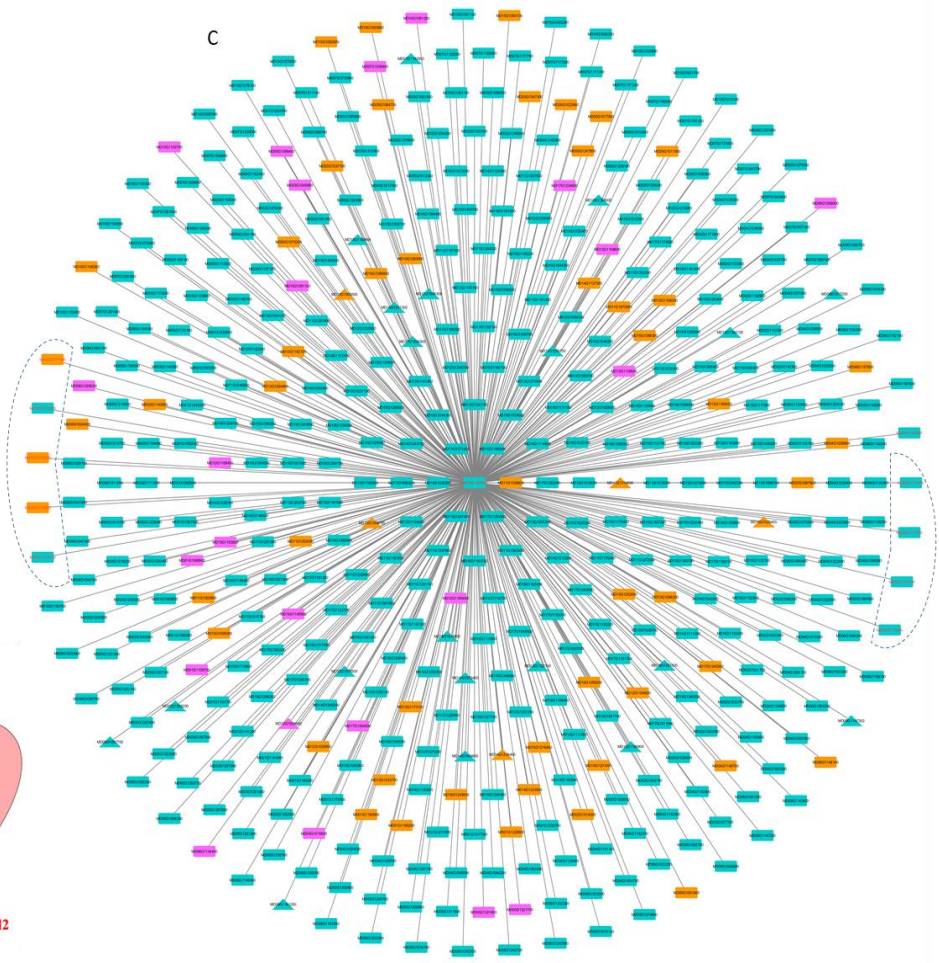


Figure 4.6. Weighted gene co-expression network analysis (WGCNA) of DEGs among progenies of phenotypes columnar, Std1 and Std2. (**A**) Correlations between WGCNA modules and tree growth habit (columnar, Std1 and Std2). (**B**) Venn Diagram analysis of the 741 member genes of WGCNA module2 associated with tree growth habit. (**C**) Primary neighbors of the *Co* gene in WGCNA module2. Orange: 57 upregulated in Std2 (in relation to columnar); Turquoise: 389 down regulated in Std2; Purple: 23 DEGs with absolute fold change ≤1.5. Triangle: transcriptional factors. DEGs in the *c2* and *c3* regions are indicated in an irregular shape on the left and right, respectively. Note that other edges are not shown.

Enriched MapMan Bins in the *Co* guided WGCNA module

Gene enrichment analysis of the member genes in module2 identified 12 MapMan bins that were over-represented (**Table 4.2**), which cover 136 of the 741 DEGs (**Table S4.7**). Among the 12 MapMan bins, M26.9 (misc.glutathione S transferases) and M26.10 (misc.cytochrome P450) are enriched most significantly. Interestingly, five of the 12 DEGs under *c2* and *c3* were

found in the 12 MapMan bins. They include MD10G1172300 encoding a glutathione S-transferase (GST) TAU 8 like protein in M26.9, and MD09G1174600, MD09G1174700, MD09G1175100, and MD10G1171100 encoding a GDSL lipase in M26.28 (misc.GDSL-motif lipase). Surprisingly, the *Co* gene (MD10G1185400), which controls columnar, is also a member in the enriched M16.8.3 (secondary metabolism.flavonoids.dihydroflavonols). These data suggest that the metabolism of dihydroflavonols and the activities of GSTs, GDSL lipases, and cytochrome P450 proteins are likely of a critical role in tree growth habit.

Table 4.2. Gene enrichment analyses of WGCNA module2 member genes in MapMan Bins.

Mapman Bins	Description	No. of DEGs expected	No. of DEGs observed	FDR p
M26.9	misc.glutathione S transferases	1	18	1.30E-11
M26.10	misc.cytochrome P450	6	29	2.22E-09
M30.2.15	signalling.receptor kinases.thaumatin like	1	13	1.09E-06
M16.8.3	secondary metabolism.flavonoids.dihydroflavonols	1	11	1.14E-05
M26.28	misc.GDSL-motif lipase	2	13	7.27E-05
M16.8.2.1	secondary metabolism.flavonoids.chalcones.naringenin-chalcone synthase	0	5	2.97E-04
M26.6	misc.O-methyl transferases	1	9	0.0019
M26.12	misc.peroxidases	2	10	0.0076
M21.2	redox.ascorbate and glutathione	1	8	0.0101
M26.19	misc.plastocyanin-like	1	7	0.0103
M26.8	misc.nitrilases, *nitrile lyases, berberine bridge enzymes, reticuline oxidases, troponine reductases	2	9	0.0226
M5.10	fermentation.aldehyde dehydrogenase	0	4	0.0355
Sum		18	136	

Discussion

Mapping of recessive traits by pooled genome sequencing in *Malus*

In an effort to adapt the pooled genome sequencing based approach for mapping a dominant trait ‘weeping’ in *Malus*, three segregation types were identified as informative and useful (Dougherty et al., 2018). They include the commonly used <**l**m × mm> (type I) in weeping pool-specific variants, and hidden <**l**m × **ll**> (type II) and <**h**k × **h**k> (type III) in variants common to both weeping and standard pools (**Table S4.9**). The first allele is designated

weeping-linked and the alleles in bold represent a DNA variant in relation to the apple reference genome (Dougherty et al., 2018). However, five segregation types <**h**k × **h**k> (A), <**l**m × **ll**> (B), <**nn** × **n**p> (C), <**l**m × mm> (D), and <pp × **n**p> (E) were inferred as informative for mapping a recessive trait when the approach was adapted in this study (**Figure S4.3**, **Table S4.3**). (Here the alleles in each first and third positions are assumed in linkage with the recessive Std2 alleles in the seed and pollen parents, respectively, and the letters in bold stand for DNA variants.) Since variants under segregation types A-C fall into group Ho-Std2/He-Col and types D-E into groups He-Std2/He-Col (**Table S4.3**), only the variants common to both pools are useful. This implicates that variants specific to pool columnar or Std2 are non-informative, contrasting to the type I variants inferred for dominant traits (Dougherty et al., 2018). Among the five segregation types, type A <**h**k × **h**k> variants are exploited commonly for mapping recessive traits while the types B-E are hidden (**Figure S4.3** and **Table S4.3**, **Table S4.9**). As to their utilities in mapping recessive traits, segregation types A-C are clearly useful as their variants are exclusively responsible for mapping *c2* and *c3* (**Figure 4.2 B-C**).

The variants of segregation types <**l**m × mm> (D), and <pp × **n**p> (E) were considered useful. However, their applicability could not be confirmed in this study, suggesting variants of types D and E may not be suitable for mapping recessive traits sometime. Viewing how variants of segregation types B and C were identified, the poor applicability of segregation types D and E might have been caused by the high number of variants (1,636,085) present in group He-Std2/He-Col (**Table S4.3**) and the filter AF 35% to 65% used in pool Std2 (**Table S4.4**). In particular, the filter is of inherent limitations as it targets variants of AF 50%, a variant allele frequency that also could be expected from many variants of unwanted segregation types, such as <**hh** × kk> and <hh × **kk**>. Indeed, the filter AF≥85%, which targets homozygous variants in

pool Std2 for segregation types A-C, is much more specific and restrictive, as there were only 70,522 variants in group Ho-Std2/He-Col (**Table S4.3**). It is recommended that segregation types A-C be the first choices for mapping recessive traits in apple and other species of heterozygous genome.

Mapping the columnar recessive repressors by mapping their dominant alleles was attempted with two ways to test if recessive model-based approaches are necessary. The first was to use directly the dominant models as described previously (Dougherty et al., 2018). The second was to use variants of segregation types <h**k** × h**k**>, <l**m** × ll>, and <nn × n**p**>, variants of which are inferred as specific to pool columnar (**Table S4.3**). However, both approaches failed to map *C2* although a 4(15.0-19.0)-Mb region overlapping with *c3* on chromosome 9 was picked-up among several others that could not be confirmed (data not shown). Inspecting the 4-Mb region on chromosome 9 showed that SNVs were markedly reduced in pool Std2 while increased in pool columnar, suggesting that the partial success in mapping of *C3* indeed represents the mapping of *c3*. Nevertheless, this success requires a low degree of heterozygosity in the *c3* region between the reference genome and pool Std2. Given the highly heterozygous nature of the apple genome and the failure to map *c2* using dominance models, it is necessary to use recessive model-based approach to map recessive traits.

Homozygous recessive loci in apple, unique to crop plants of heterozygous genome?

This study identified *c2* and *c3* as homozygous genetic loci controlling recessive trait Std2. Despite the high density of homozygous SNVs (AF≥85%) of segregation types A-C, the DNA sequences in the *c2* and *c3* genomic regions in pool Std2 were far from being identical. In the *c2* peak region of 2(25.0-27.0)-Mb, 4710 variants of segregation types A-C were identified

(**Figure 4.2 B-C**), accounting for 52.9% of the total SNVs (8912) common to both pools, i.e. 47.1% were heterozygous SNVs. Similarly, in the *c3* peak region of 2(14.0-16.0)-Mb, 2160 types A-C variants were identified (**Figure 4.2 B-C**), accounting for only 46.7% of the total common SNVs (4628). These data suggested that many SNVs remain heterozygous in the *c2* and *c3* regions in pool Std2. This raises the possibility if *c2* and *c3* are recessive compound heterozygous loci that are reported commonly in human and animals (Takaku et al., 1998; Zhao et al., 2006; Zhong et al., 2017), which describe a gene locus of two different recessive mutant alleles that confers a recessive condition or disease. However, the high density of homozygous variants of segregation types A-C in the *c2* and *c3* regions were also present in the coding regions of many genes (**Figure S4.8, Figure S4.9**). Therefore, it is more likely that the homozygous recessive inheritance of the Std2 trait was determined by the underlying genes carrying homozygous DNA variants. Such recessive loci that are determined by genes of homozygous SNVs in heterozygous genomic regions may reflect an important distinction of apple from the recessive homozygous loci in inbreeding crops, such as rice and tomato, and from the recessive compound heterozygous loci in human and animal (Takaku et al., 1998; Zhao et al., 2006; Zhong et al., 2017).

Effect of the *c2* and *c3* interactions on columnar repression with "incomplete penetrance"

The observed Std2 frequencies in the three years (2009, 2011 and 2015) were high in double recessive genotype *c2c2 c3c3* (0.867, 0.667 and 0.448), medium in single-recessive carriers *c2c2 c3C3* (0.478, 0.348 and 0.136) and *c2C2 c3c3* and *C2C2 c3c3* (0.237, 0.186 and 0.153), and low in non-recessive carriers *c2C2 c3C3* and *C2C2 c3C3* (0.107, 0.089 and 0.018) (**Figure 4.4**). Since the frequency of Std2 in the double recessive carrier was even more than the

combined fractions of *c2c2* and *c3c3* in single recessive carriers in the three years (0.715, 0.534, and 0.289), the two loci were proposed to repress columnar through additive gene interactions. The hypothesis is that the homozygous recessive genotypes *c2c2* and *c3c3* each would drive a certain fraction of the single recessive carriers to express Std2 at a given year while the double recessive genotype *c2c2 c3c3* would empower a higher fraction or all of its carriers to express Std2. Overall, this proposal explains the data well although the small fraction of Std2 in non-recessive carries could not be accounted for. Apparently, the Std2 frequencies in the double recessive carrier *c2c2 c3c3* were lower than 100% in the three years, suggesting the additive effect of *c2* and *c3* could drive only "incomplete penetrance" of Std2 that could be reduced to a lower penetrance by tree age.

Incomplete penetrance and variable expressivity have been documented well in plant (Mazzucato et al., 2015; Sekhon & Chopra, 2009), animal (Eichers et al., 2006; Raj, Rifkin, Andersen, & van Oudenaarden, 2010) and human (Bourgeois et al., 1998; Giudicessi & Ackerman, 2013). Depending upon studies, the range of incomplete penetrance varied widely. For example, the range of penetrance for human long QT syndrome (LQTS) in individual LQTS families were anywhere between 25% and 100% (Giudicessi & Ackerman, 2013), whereas the penetrance of aberrations in cotyledon morphology and carpelloid stamens in homozygous siblings (BC_1F_2) from an *Aux/IAA9* frameshift mutation in tomato were reported with 47.1% and 41.0%, respectively (Mazzucato et al., 2015). In addition, age-dependent penetrance and expressivity of certain phenotype appeared to be common in animal (Eichers et al., 2006) and plant (Ashri, 1970) as well. The causal factors for the phenomenon of incomplete penetrance have been attributed to environments, interactions with other genes, and epigenetic regulation of expression of the underlying genes (Lalucque & Silar, 2004; Raj et al., 2010; Wittmeyer et al.,

2018). Since the retroposon induced high expression of the *Co* gene in columnar (MD10G1185400) (Okada et al., 2016; Otto et al., 2014; Wolters et al., 2013) is drastically repressed in Std2 (**Figure 4.5 E**, **Figure S4.5 A**), it is possible that *c2* and *c3* would interact with *Co* and/or involve an epigenetic mechanism that regulates the expression of *Co*, thereby the penetrance of phenotype Std2.

Candidate genes under *c2* and *c3*

The DEGs between columnar and Std2 under *c2* and *c3* (**Figure 4.5 D**) are considered an important group of candidate genes, of which the four GDSL lipase encoding genes and the GST encoding gene (MD10G1172300) are of particular interest. MD09G1174600, MD09G1174700, MD09G1175100 and MD10G1171100 (the four GDSL lipase encoding genes) directly contributed to the enrichment of M26.28 (misc.GDSL-motif lipase) in the WGCNA module2 (**Table 4.2**). Under *c3*, the first three GDSLs were all downregulated in Std1 and Std2 and were expressed at relatively lower levels, similar to the *Co* gene (**Figure 4.5 D**-**E**). The *Arabidopsis* counterpart of MD09G1174600 is At1g53940 (GLIP2, AtGELP20), and that of MD09G1174700 and MD09G1175100 is At5g40990 (GLIP1, AtGELP97). Interestingly, GLIP1 and GLIP2 are most closely related member genes of Clade IIIa in the GDSL lipase gene family in *Arabidopsis* (Lai, Huang, Chen, Chan, & Shaw, 2017), suggesting the three GDSLs form a closely related gene cluster under *c3*. T-DNA knockout lines of At1g53940 (GLIP2) and At5g40990 (GLIP1) were similarly more sensitive to pathogen *E. carotovora* than their wild type controls (Lai et al., 2017; Lee, Kim, Kwon, Jin, & Park, 2009), implicating their roles in plant response to biotic stress. However, the T-DNA knockout lines of At1g53940 (GLIP2) were observed also with drastically increased lateral roots, impaired gravitropic response of shoots, and increased levels

of the transcripts of *IAA1* and *IAA2*, indicating At1g53940 (GLIP2) negatively regulates auxin signaling (Lee et al., 2009), which is important in plant growth and development.

MD10G1171100 under *c2* showed an opposite expression profile of the three GDSLs discussed above. The *Arabidopsis* counterpart of MD10G1171100 is At4G28780 (AtGELP82), which is a member in Clade IIb of the GDSL lipase gene family (Lai et al., 2017). This clade includes a well-characterized gene, At5g33370 (AtGELP95) that encodes CUTIN SYNTHASE2 (CUS2), which is essential for the development of cuticular ridges in *Arabidopsis* sepals (Hong, Brown, Segerson, Rose, & Roeder, 2017). *CUS2* is mostly expressed in various organs in reproductive stage while At4G28780 is expressed in many tissues in both vegetative and reproductive stages (Schmid et al., 2005), implicating a complex role of the Clade IIb genes in plant development.

MD10G1172300, encoding a glutathione S transferase (GST) under *c2*, is a member in the most significantly enriched MapMan Bin M26.9 (misc.glutathione S transferases) in the WGCNA module2. MD10G1172300 is the apple counterpart of *Arabidopsis* GSTU8 (At3g09270), one of the 28 Tau (U) class GSTs in the GST gene family (Dixon & Edwards, 2010). Although the function of GSTU8 is not clear, several members of GSTUs, such as GSTU17, GSTU19 and GSTU20 have been shown to regulate plant photomorphogenesis and/or root development (Chen et al., 2007; Gallé et al., 2019; Horváth et al., 2019; Jiang et al., 2010).

In addition to the DEGs, genes that did not expressed differentially between columnar and Std2 in statistics under *c2* and *c3* cannot be ruled out as candidate genes. These genes may include transcription regulators (MD10G1165100, MD09G1170000, MD09G1170200, and MD09G1170800), transcription factors (MD10G1159600, MD10G1159800, MD10G1163600, and MD10G1170600, MD09G1174400, and MD09G1175700), Sterile alpha motif (SAM)

domain-containing protein encoding gene (MD10G1161300), genes related to phototropic-response (MD10G1164500) and auxin response (MD10G1160000, MD10G1176400), and others. Clearly, dedicated studies are needed to determine if any of these candidate genes discussed above or others are the casual genes underlying *c2* and *c3* that repress the *Co* gene expression and columnar phenotype.

Conclusions

By exploring DNA variant segregation types in pooled genome sequencing, this study elucidated the genetic basis on which SNVs of segregation types A-E can be employed together with the AFDDD mapping strategy to map recessive traits in apple. Application of the mapping strategy successfully identified two recessive repressors *c2* and *c3* associated with columnar repression, which are located on chromosomes 10 and 9, respectively. An important mechanism through which *c2* and *c3* mediate the columnar repression is to repress the *Co* gene expression. The identification of the *Co* gene-guided WGCNA module offers further clues on how the causal genes underlying *c2* and *c3* may function to repress the *Co* gene expression. Overall, this study demonstrates an effective approach for mapping recessive traits in apple and other out-crossing crop species and provides new insights into genetic and molecular regulation of columnar growth habit in apple.

Acknowledgments

This work was financially supported by a grant award (IOS-1339211) from NSF-Plant Genome Research Program.

REFERENCES

Ashri, A. (1970). A dominant mutation with variable penetrance and expressivity induced by diethyl sulfate in peanuts, *Arachis hypogaea* L. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 9*(5), 473-480.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., & Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics, 24*(2), 282-284.

Bai, T., Zhu, Y., Fernández-Fernández, F., Keulemans, J., Brown, S., & Xu, K. (2012). Fine genetic mapping of the Co locus controlling columnar growth habit in apple. *Molecular Genetics and Genomics, 287*, 437-450.

Bai, Y., Dougherty, L., Cheng, L., Zhong, G.-Y., & Xu, K. (2015). Uncovering co-expression gene network modules regulating fruit acidity in diverse apples. *BMC Genomics, 16:612* 1-16.

Bai, Y., Dougherty, L., & Xu, K. (2014). Towards an improved apple reference transcriptome using RNA-seq. *Mol Genet Genomics, 289*(3), 427-438.

Baldi, P., Wolters, P., Komjanc, M., Viola, R., Velasco, R., & Salvi, S. (2013). Genetic and physical characterisation of the locus controlling columnar habit in apple (Malus × domestica Borkh.). *Molecular Breeding, 31*(2), 429-440.

Blazek, J. (2013). Performance of tree growth characteristics in selected progenies of columnar apple cultivars. *Acta Hortic., 976*, 345-353.

Bourgeois, P., Bolcato-Bellemin, A.-L., Danse, J.-M., Bloch-Zupan, A., Yoshiba, K., Stoetzel, C., & Perrin-Schmitt, F. (1998). The Variable Expressivity and Incomplete Penetrance of the twist-Null Heterozygous Mouse Phenotype Resemble Those of Human Saethre-Chotzen Syndrome. *Human Molecular Genetics, 7*(6), 945-957.

Chen, I. C., Huang, I. C., Liu, M. J., Wang, Z. G., Chung, S. S., & Hsieh, H. L. (2007). Glutathione S-transferase interacting with far-red insensitive 219 is involved in phytochrome A-mediated signaling in Arabidopsis. *Plant Physiol, 143*(3), 1189-1202.

Conner, P. J., Brown, S. K., & Weeden, N. F. (1997). Randomly amplified polymorphic DNA-based genetic linkage maps of three apple cultivars. *Journal of the American Society for Horticultural Science, 122*(3), 350-359.

Conner, P. J., Brown, S. K., & Weeden, N. F. (1998). Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. *Theoretical and Applied Genetics, 96*(8), 1027-1035.

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., Di Pierro, E. A., Gouzy, J., Rees, D. J. G., Guerif, P., Muranty, H., Durel, C.-E., Laurens, F., Lespinasse, Y., Gaillard, S., Aubourg, S., Quesneville, H., Weigel, D., van de Weg, E., Troggio, M., & Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat Genet, 49*, 1099-1106.

Dixon, D. P., & Edwards, R. (2010). *Glutathione Transferases* (Vol. 2010): BIOONE.

Dougherty, L., Singh, R., Brown, S., Dardick, C., & Xu, K. (2018). Exploring DNA variant segregation types in pooled genome sequencing enables effective mapping of weeping trait in Malus. *Journal of Experimental Botany, 69*(7), 1499-1516.

Eichers, E. R., Abd-El-Barr, M. M., Paylor, R., Lewis, R. A., Bi, W., Lin, X., Meehan, T. P., Stockton, D. W., Wu, S. M., Lindsay, E., Justice, M. J., Beales, P. L., Katsanis, N., & Lupski, J. R. (2006). Phenotypic characterization of Bbs4 null mice reveals age-dependent penetrance and variable expressivity. *Human Genetics, 120*(2), 211-226.

Evenson, R. E., & Gollin, D. (2003). Assessing the Impact of the Green Revolution, 1960 to 2000. *Science, 300*(5620), 758-762.

Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics, 23*(2), 257-258.

Fernandez-Fernandez, F., Evans, K. M., Clarke, J. B., Govan, C. L., James, C. M., Maric, S., & Tobutt, K. R. (2008). Development of an STS map of an interspecific progeny of Malus. *Tree Genetics & Genomes, 4*(3), 469-479.

Gallé, Á., Czékus, Z., Bela, K., Horváth, E., Ördög, A., Csiszár, J., & Poór, P. (2019). Plant Glutathione Transferases and Light. *Frontiers in plant science, 9*(1944).

Giudicessi, J. R., & Ackerman, M. J. (2013). Determinants of incomplete penetrance and variable expressivity in heritable cardiac arrhythmia syndromes. *Translational Research, 161*(1), 1-14.

Hemmat, M., Weeden, N. F., Conner, P. J., & Brown, S. K. (1997). A DNA marker for columnar growth habit in apple contains a simple sequence repeat. *Journal of the American Society for Horticultural Science, 122*(3), 347-349.

Hong, L., Brown, J., Segerson, N. A., Rose, J. K. C., & Roeder, A. H. K. (2017). CUTIN SYNTHASE 2 Maintains Progressively Developing Cuticular Ridges in Arabidopsis Sepals. *Molecular Plant, 10*(4), 560-574.

Horváth, E., Bela, K., Holinka, B., Riyazuddin, R., Gallé, Á., Hajnal, Á., Hurton, Á., Fehér, A., & Csiszár, J. (2019). The Arabidopsis glutathione transferases, AtGSTF8 and AtGSTU19 are involved in the maintenance of root redox homeostasis affecting meristem size and salt stress sensitivity. *Plant Science, 283*, 366-374.

Jiang, H.-W., Liu, M.-J., Chen, I.-C., Huang, C.-H., Chao, L.-Y., & Hsieh, H.-L. (2010). A Glutathione *S*-Transferase Regulated by Light and Hormones Participates in the Modulation of Arabidopsis Seedling Development. *Plant Physiology, 154*(4), 1646-1658.

Kenis, K., & Keulemans, J. (2007). Study of tree architecture of apple (Malus x domestica Borkh.) by QTL analysis of growth traits. *Molecular Breeding, 19*(3), 193-208.

Kim, M. Y., Song, K. J., Hwang, J. H., Shin, Y. U., & Lee, H. J. (2003). Development of RAPD and SCAR markers linked to the Co gene conferring columnar growth habit in apple (Malus pumila Mill.). *Journal of Horticultural Science & Biotechnology, 78*(4), 512-517.

Lai, C.-P., Huang, L.-M., Chen, L.-F. O., Chan, M.-T., & Shaw, J.-F. (2017). Genome-wide analysis of GDSL-type esterases/lipases in Arabidopsis. *Plant Molecular Biology, 95*(1), 181-197.

Lalucque, H., & Silar, P. (2004). Incomplete Penetrance and Variable Expressivity of a Growth Defect as a Consequence of Knocking Out Two K^+ Transporters in the Euascomycete Fungus *Podospora anserina*. *Genetics, 166*(1), 125-133.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *Bmc Bioinformatics, 9*(1), 559.

Lapins, K. O. (1969). Segregation of compact growth types in certain apple seedling progenies. *Canadian Journal of Plant Science, 49*, 765-768.

Lapins, K. O. (1976). Inheritance of compact growth type in apple. *Journal of the American Society for Horticultural Science, 101*(2), 133-135.

Lauri, P. E., & Lespinnasse, J. M. (1993). The relationship between cultivar fruiting-type and fruiting branch characteristics in apple trees. *Acta Hortic, 349*, 259-263.

Lee, D. S., Kim, B. K., Kwon, S. J., Jin, H. C., & Park, O. K. (2009). Arabidopsis GDSL lipase 2 plays a role in pathogen defense via negative regulation of auxin signaling. *Biochemical and Biophysical Research Communications, 379*(4), 1038-1042.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A. R., Stitt, M., & Usadel, B. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell & Environment, 37*, 1250–1258.

Mazzucato, A., Cellini, F., Bouzayen, M., Zouine, M., Mila, I., Minoia, S., Petrozza, A., Picarella, M. E., Ruiu, F., & Carriero, F. (2015). A TILLING allele of the tomato Aux/IAA9 gene offers new insights into fruit set mechanisms and perspectives for breeding seedless tomatoes. *Molecular Breeding, 35*(1), 22.

Meisel, L., Fonseca, B., González, S., Baeza-Yates, R., Cambiazo, V., Campos, R., Gonzalez, M., Orellana, A., Retamales, J., & Silva, H. (2005). A rapid and efficient method for purifying high quality total RNA from peaches (Prunus persica) for functional genomics analyses. *Biological Research, 38*(1), 83-88.

Meulenbroek, B., Verhaegh, J., & Janse, J. (1998). Inheritance studies with columnar type trees. *Acta Hort, 484*, 255-260.

Morimoto, T., & Banno, K. (2015). Genetic and physical mapping of Co, a gene controlling the columnar trait of apple. *Tree Genetics & Genomes, 11*(1), 807.

Moriya, S., Iwanami, H., Kotoda, N., Takahashi, S., Yamamoto, T., & Abe, K. (2009). Development of a Marker-assisted Selection System for Columnar Growth Habit in Apple Breeding. *Journal of the Japanese Society for Horticultural Science, 78*(3), 279-287.

Moriya, S., Okada, K., Haji, T., Yamamoto, T., & Abe, K. (2012). Fine mapping of Co, a gene controlling columnar growth habit located on apple (Malus×domestica Borkh.) linkage group 10. *Plant Breeding, 131*(5), 641-647.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods, 5*(7), 621-628.

Okada, K., Wada, M., Moriya, S., Katayose, Y., Fujisawa, H., Wu, J., Kanamori, H., Kurita, K., Sasaki, H., Fujii, H., Terakami, S., Iwanami, H., Yamamoto, T., & Abe, K. (2016). Expression of a putative dioxygenase gene adjacent to an insertion mutation is involved in the short internodes of columnar apples (Malus × domestica). *Journal of Plant Research, 129*(6), 1109-1126.

Otto, D., Petersen, R., Brauksiepe, B., Braun, P., & Schmidt, E. (2014). The columnar mutation ("Co gene") of apple (Malus × domestica) is associated with an integration of a Gypsy-like retrotransposon. *Molecular Breeding, 33*, 863-880.

Petersen, R., & Krost, C. (2013). Tracing a key player in the regulation of plant architecture: the columnar growth habit of apple trees (Malus × domestica). *Planta, 238*(1), 1-22.

Raj, A., Rifkin, S. A., Andersen, E., & van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *Nature, 463*, 913.

Robinson, T., Hoying, S., Sazo, M. M., DeMarree, A., & Dominguez, L. (2013). A vision for apple orchard systems of the future. *NY Fruit Q, 21*, 11-16.

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., Pico, A. R., Bader, G. D., & Ideker, T. (2012). A travel guide to Cytoscape plugins. *Nat Meth, 9*(11), 1069-1076.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., & Lohmann, J. U. (2005). A gene expression map of Arabidopsis thaliana development. *Nature Genetics, 37*(5), 501-506.

Sekhon, R. S., & Chopra, S. (2009). Progressive Loss of DNA Methylation Releases Epigenetic Gene Silencing From a Tandemly Repeated Maize Myb Gene. *Genetics, 181*(1), 81-91.

Takaku, K., Oshima, M., Miyoshi, H., Matsui, M., Seldin, M. F., & Taketo, M. M. (1998). Intestinal Tumorigenesis in Compound Mutant Mice of both Dpc4(Smad4) and Apc Genes. *Cell, 92*(5), 645-656.

Taylor, M., & Granatstein, D. (2013). A Cost Comparison of Organic and Conventional Apple Production in the State of Washington. *Crop Management, 12*(1).

Tian, Y.-K., Wang, C.-H., Zhang, J.-S., James, C., & Dai, H.-Y. (2005). Mapping Co, a gene controlling the columnar phenotype of apple, with molecular markers. *Euphytica, 145*(1), 181-188.

Tobutt, K. R. (1984). Breeding Columnar Apple Varieties at East Malling. *Scientific Horticulture, 35*, 72-77.

Vávra, R., BlaÅ¾ek, J., Vejl, P., & JonÃ¡Å¡, M. (2017). Evaluation of biennial bearing of apple genotypes with columnar tree growth habit. *Acta Hortic, 1172*, 395-398.

Wada, M., Iwanami, H., Moriya, S., Hanada, T., Moriya-Tanaka, Y., Honda, C., Shimizu, T., Abe, K., & Okada, K. (2018). A root-localized gene in normal apples is ectopically expressed in aerial parts of columnar apples. *Plant Growth Regulation, 85*(3), 389–398.

Wang, A., Aldwinckle, H., Forsline, P., Main, D., Fazio, G., Brown, S., & Xu, K. (2012). EST contig-based SSR linkage maps for Malus × domestica cv Royal Gala and an apple scab

resistant accession of M. sieversii, the progenitor species of domestic apple. *Molecular Breeding, 29*, 379-397.

West, T., Sullivan, R., Seavert, C., & Castagnoli, S. (2012). Enterprise Budget Apples, Medium Density, North Central Region. In (pp. 5): Oregon State University.

Wittmeyer, K., Cui, J., Chatterjee, D., Lee, T.-f., Tan, Q., Xue, W., Jiao, Y., Wang, P.-H., Gaffoor, I., Ware, D., Meyers, B. C., & Chopra, S. (2018). The Dominant and Poorly Penetrant Phenotypes of Maize *Unstable factor for orange1* Are Caused by DNA Methylation Changes at a Linked Transposon. *The Plant Cell, 30*(12), 3006-3023.

Wolters, P. J., Schouten, H. J., Velasco, R., Si-Ammour, A., & Baldi, P. (2013). Evidence for regulation of columnar habit in apple by a putative 2OG-Fe(II) oxygenase. *New Phytologist, 200*(4), 993-999.

Xu, K., Wang, A., & Brown, S. (2012). Genetic characterization of the Ma locus with pH and titratable acidity in apple. *Molecular Breeding, 30*(2), 899–912.

Zhao, Z., Tuakli-Wosornu, Y., Lagace, T. A., Kinch, L., Grishin, N. V., Horton, J. D., Cohen, J. C., & Hobbs, H. H. (2006). Molecular Characterization of Loss-of-Function Mutations in PCSK9 and Identification of a Compound Heterozygote. *The American Journal of Human Genetics, 79*(3), 514-523.

Zhong, K., Zhu, G., Jing, X., Hendriks, A. E. J., Drop, S. L. S., Ikram, M. A., Gordon, S., Zeng, C., Uitterlinden, A. G., Martin, N. G., Liu, F., & Kayser, M. (2017). Genome-wide compound heterozygote analysis highlights alleles associated with adult height in Europeans. *Human Genetics, 136*(11), 1407-1417.

Zhu, Y. D., Zhang, W., Li, G. C., & Wang, T. (2007). Evaluation of inter-simple sequence repeat analysis for mapping the Co gene in apple (Malus pumila Mill.). *Journal of Horticultural Science & Biotechnology*, *82*(3), 371-376.

Supplementary Figures

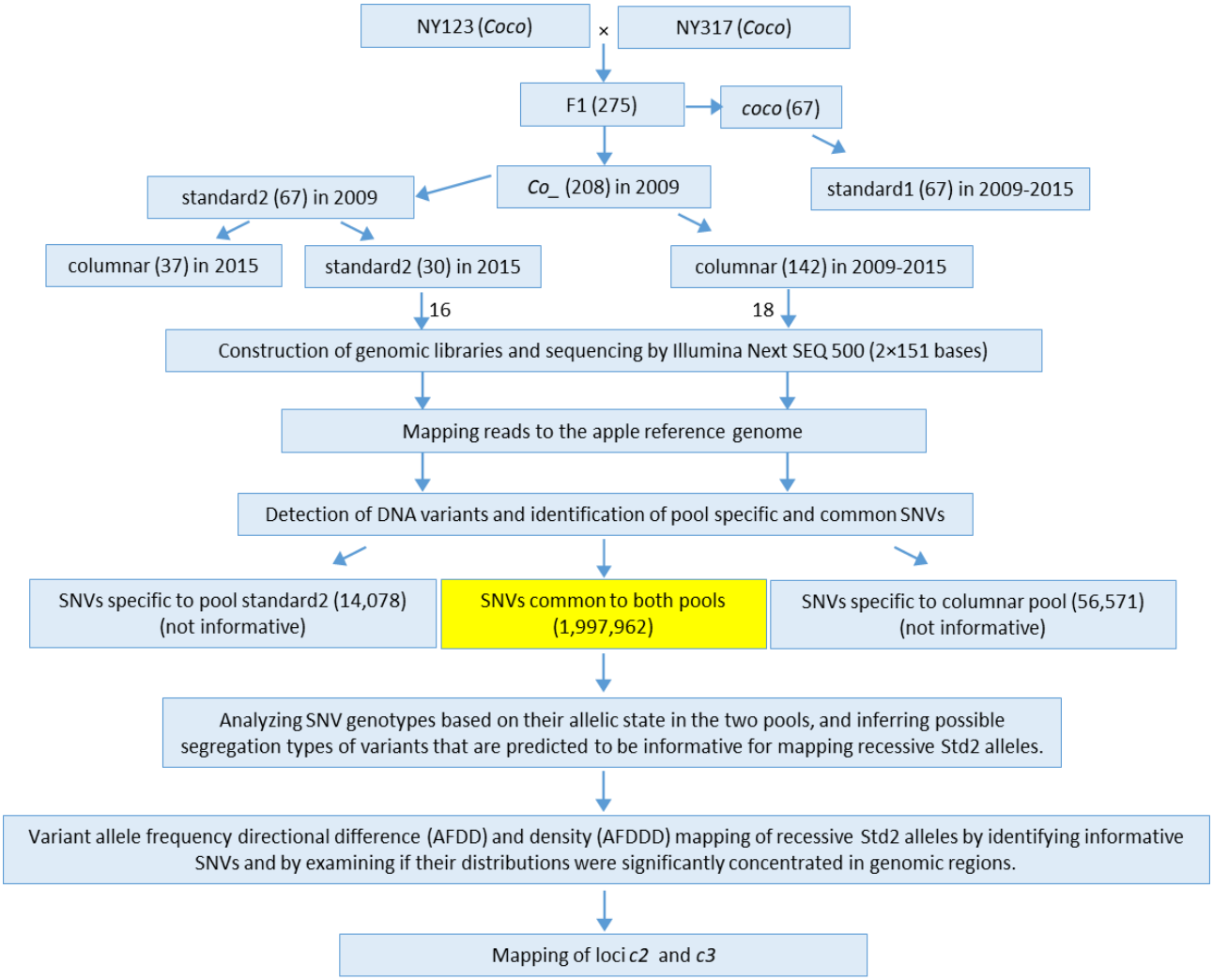


Figure S4.1. A flowchart illustrating the procedure in pooled genome sequencing and variant allele frequency directional difference (AFDD) and density (AFDDD) mapping of the recessive standard2 (Std2) phenotype. Single nucleotide variants (SNVs) were identified with the following settings: reads overage =20-200; no complex genotype and variant allele frequency (AF)≥15%. Pool specific variants were filtered further against the reads mappings in the contrasting pool in which variant AF≤10% was set.

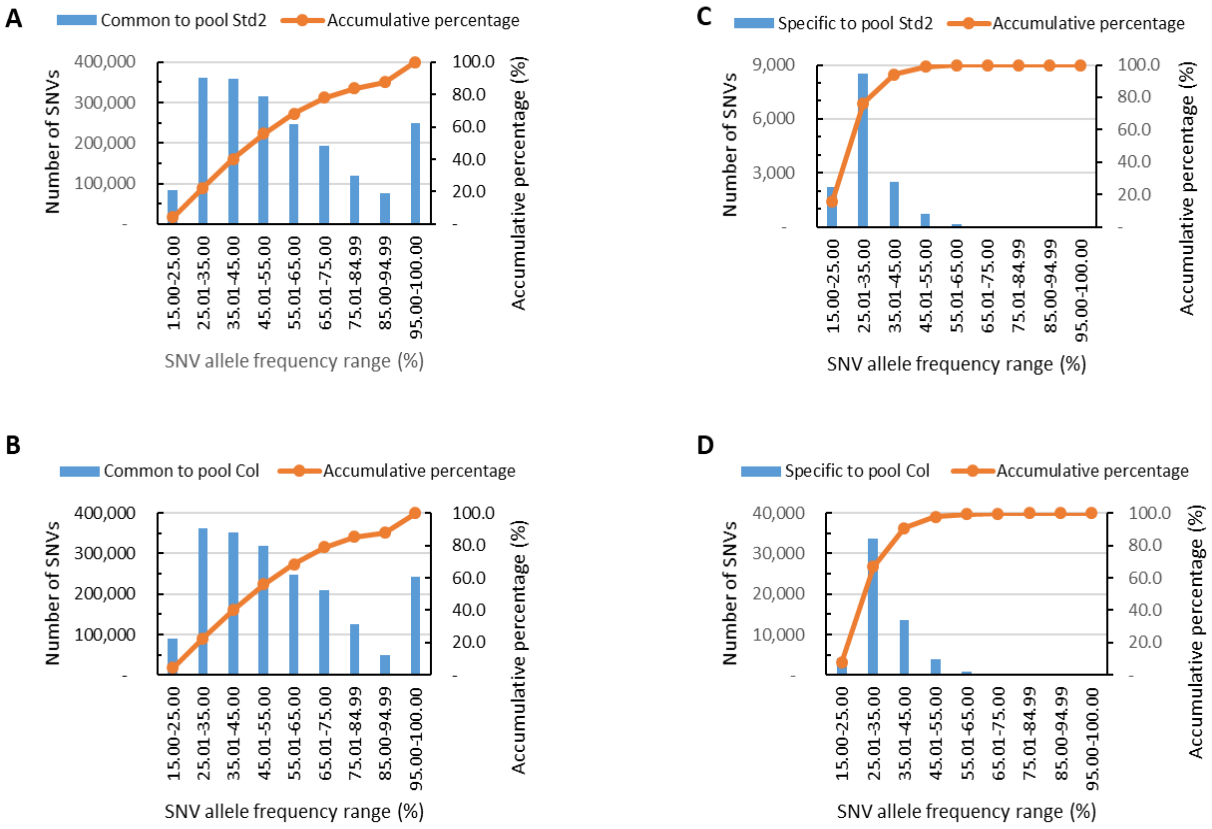


Figure S4.2. Distribution of single nucleotide variants (SNVs) under various SNV allele frequencies. (**A** and **B**) Distribution of the 1,997,962 SNVs common to both pools in pool Std2 (**A**) and columnar (**B**). (**C** and **D**) Distribution of pool specific SNVs in pool Std2 (14,078 SNVs) (**C**) and pool columnar (56,571 SNVs) (**D**).

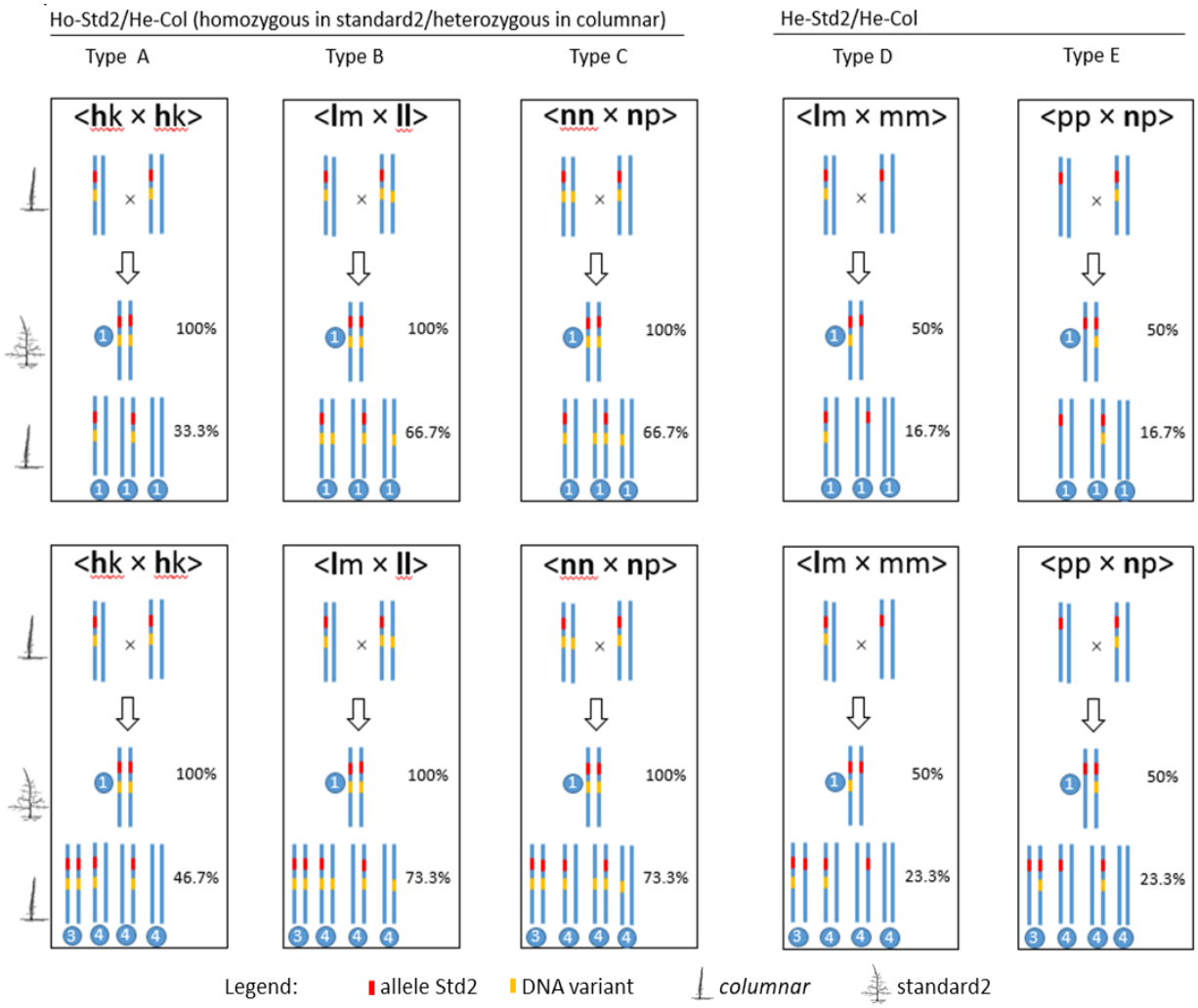


Figure S4.3. Schematic representations of informative variant segregation types (types A-E) inferred from SNV groups Ho-Std2/He-Col (homozygous in pool standard2/heterozygous in pool columnar) and He-Std2/He-Col under models of one (top panel) and two (bottom panel) recessive repressors. Non-informative segregation types were listed in Supplementary Table S3. Each segregation type is illustrated in a rectangle box that includes the two parents at the top, one representative standard2 progeny in the middle, and three or four columnar progenies at the bottom. The long vertical lines in blue represent the chromosomal segment harboring the recessive repressor locus. The red and orange short vertical lines represent the recessive repressor allele(s) and DNA variants in relation to the reference genome, respectively. The tree-drawings indicate columnar and standard2 phenotypes, respectively. The numbers within the blue circles stand for the genotype fraction number in the progeny. Total fractions under one- and two-gene model are 4 and 16, respectively. The expected allele frequencies (%) of DNA variants in pools standard2 and columnar are given accordingly. In segregation type denotation, each letter denotes one of the four DNA bases and the alleles in each first and third positions are assumed in linkage with the recessive standard2 alleles in the seed and pollen parents, respectively. Std2: standard2; Col: columnar.

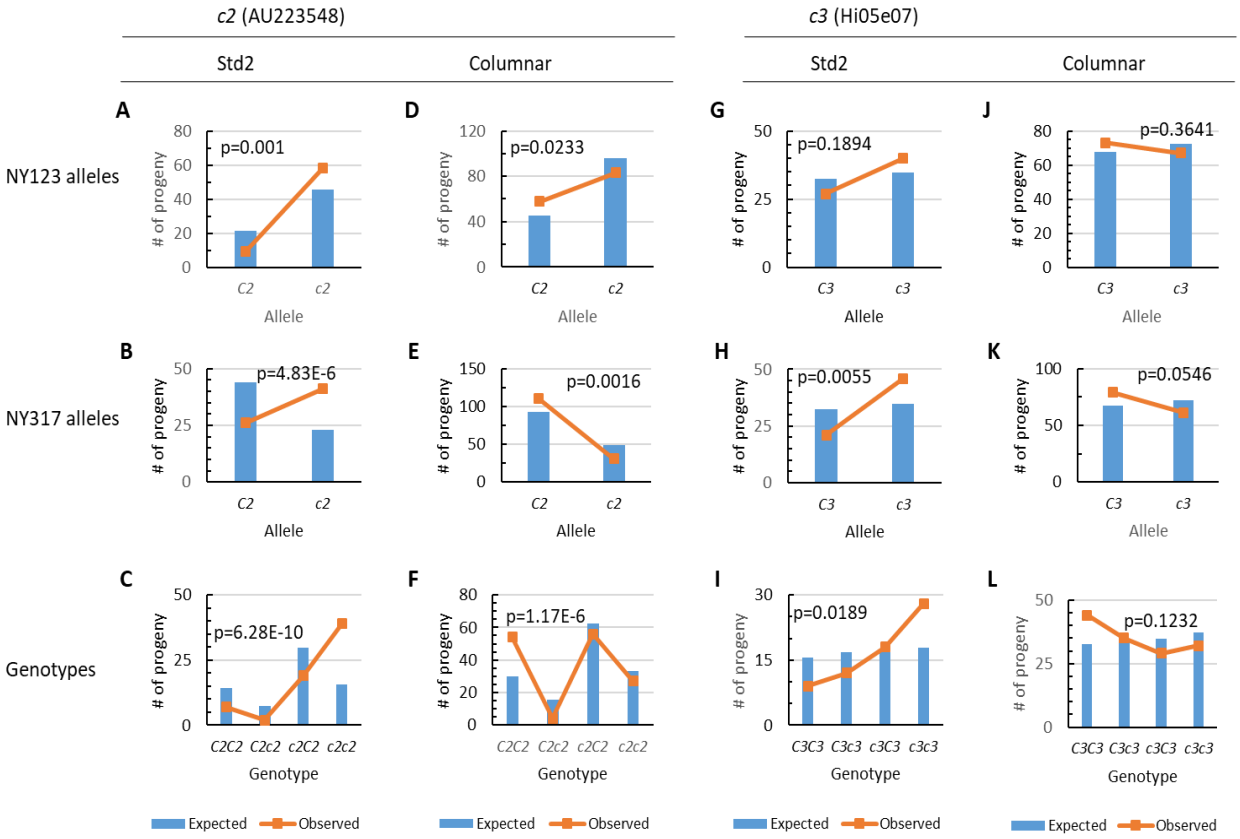


Figure S4.4. The expected and observed frequencies of alleles c2 and c3 and their genotypes in the standard2 (Std2) and columnar sub-populations in 2009. (**A-F**) The expected and observed frequencies of alleles *c2* from parents NY123 (**A**, **D**) and NY317 (**B**, **E**) in sub-populations Std2 (**A**, **B**) and columnar (**D**, **E**), and the expected and observed *c2* genotype frequencies in sub-populations Std2 (**C**) and columnar (**F**). (**G-L**) The expected and observed frequencies of alleles *c3* from parents NY123 (**G**, **J**) and NY317 (**H**, **K**) in sub-populations Std2 (**G**, **H**) and columnar (**J**, **K**), and the expected and observed *c3* genotype frequencies in sub-populations Std2 (**I**) and columnar (**L**). The p values indicate levels of significance in chi-square test. The *c2* and *c3* loci were represented and investigated with markers AU223548 and Hi05e07, respectively.

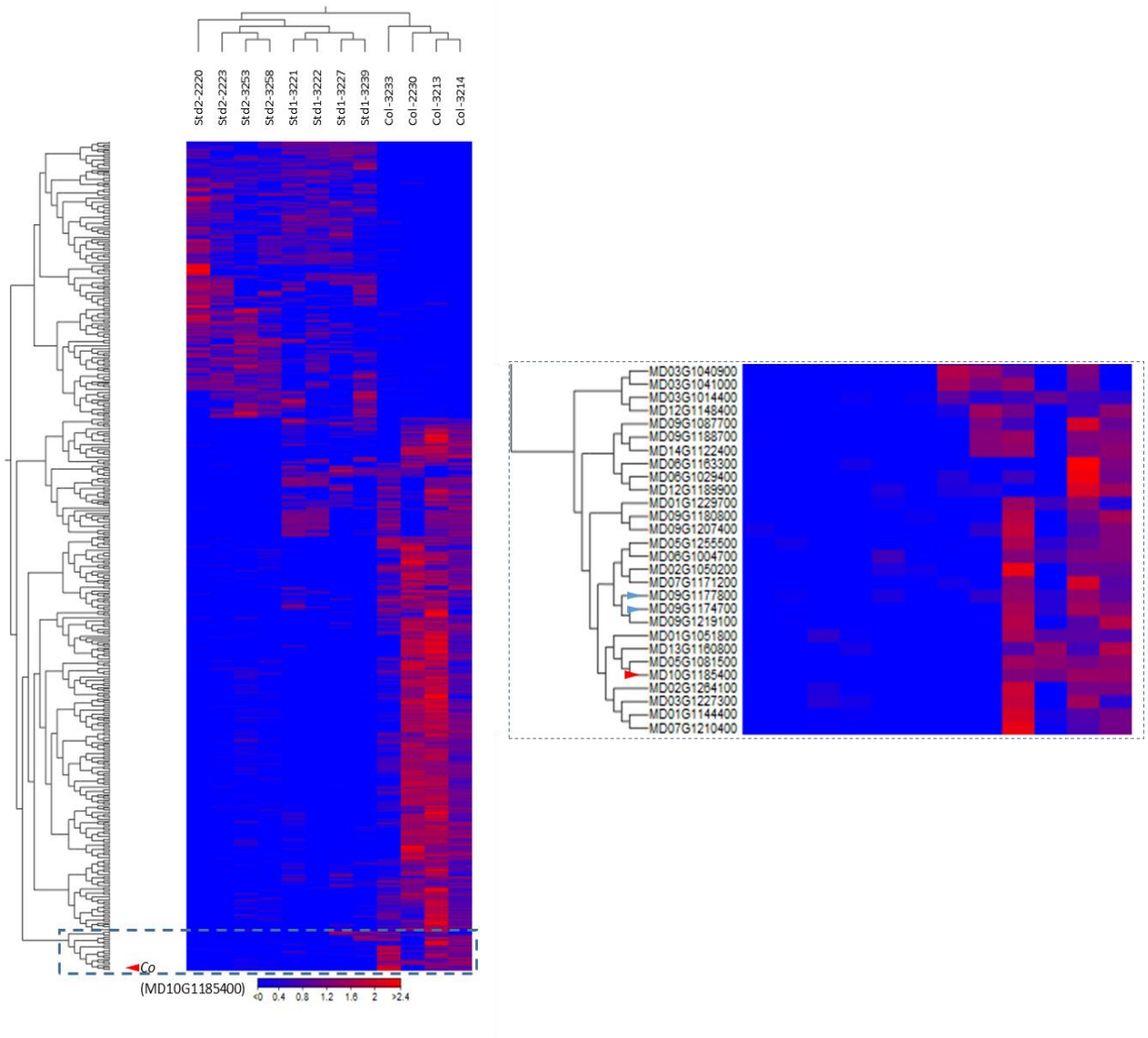


Figure S4.5. Heat map representation of DEGs (588) between columnar and standard2 progeny. There are 196 DEGs up-regulated and 392 down-regulated in Std2, respectively. The hierarchical cluster trees indicate the relationships among genes (left) and samples (top), respectively. The broken-line boxes show the section containing the Co gene (indicated by a red arrow) in the overall and zoom-in views, respectively. The arrows in blue indicate DEGs in the *c3* region.

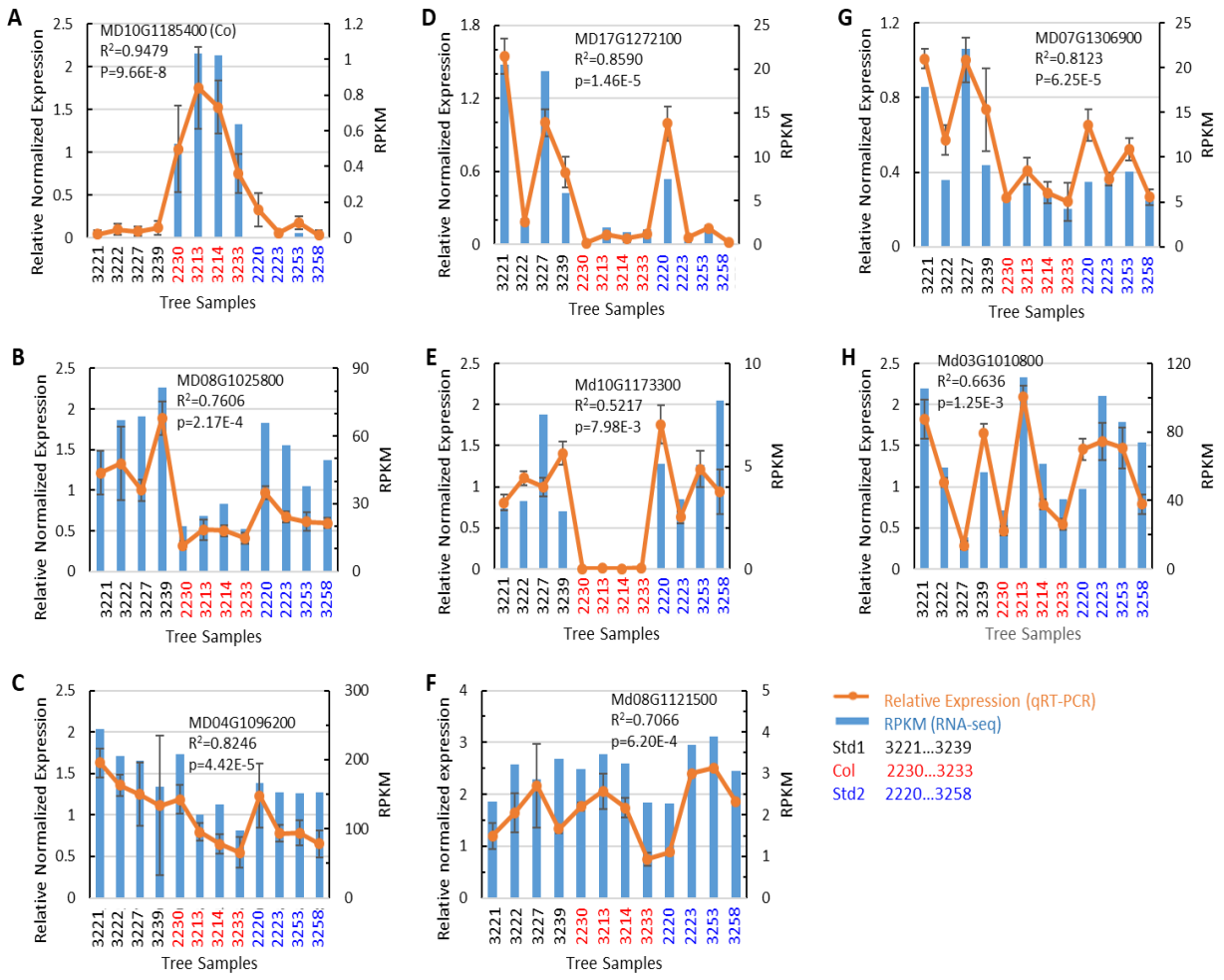


Figure S4.6. qRT-PCR validation of RNA-seq expression quantification in *Co* (MD10G1185400) and other seven genes selected randomly. The expression correlations between qRT-PCR and RNA-seq are highly significant (R^2=0.5217 to 0.9479; p=7.98E-3 to 9.66E-8; n=12).Error bars: SD.

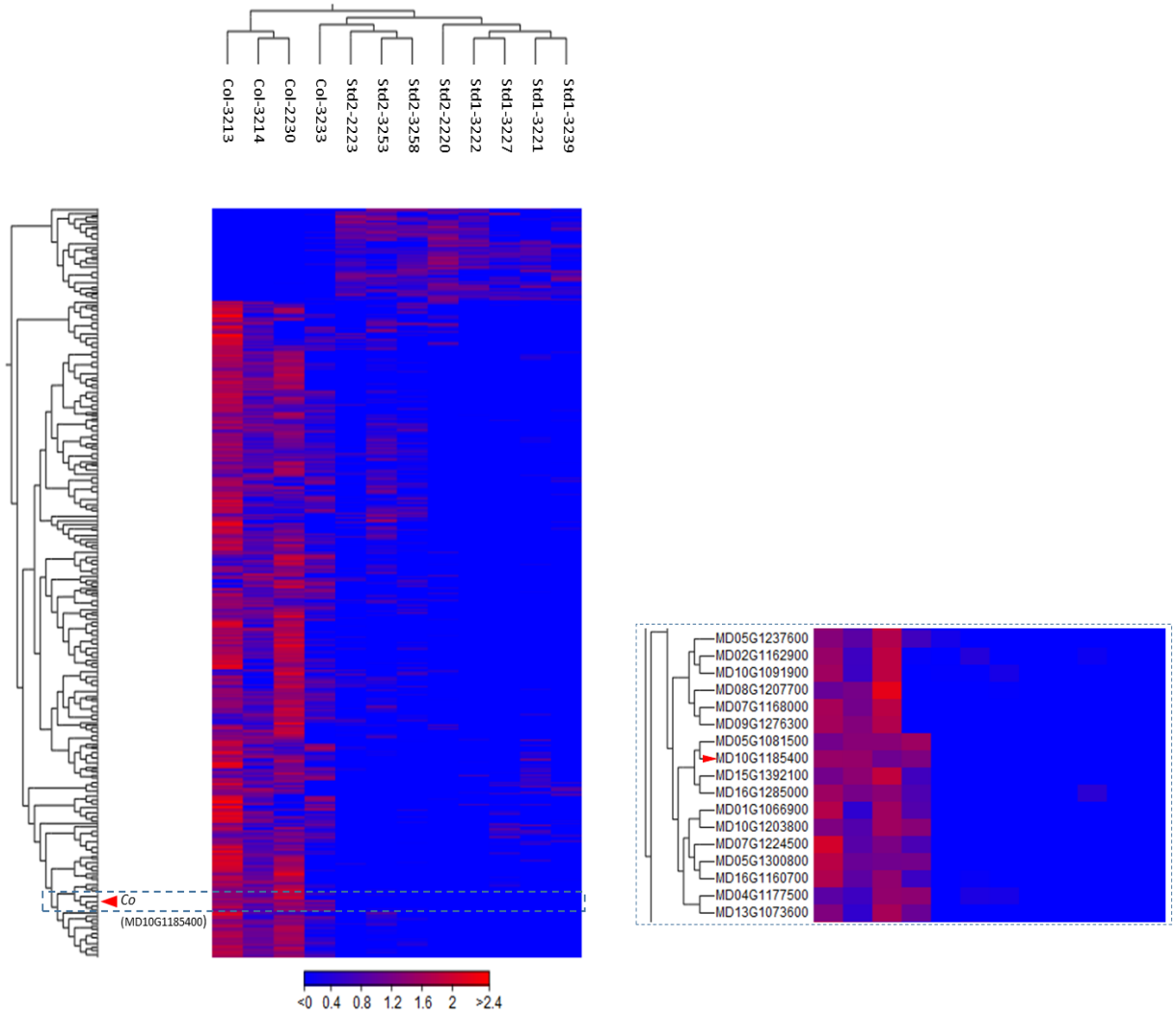


Figure S4.7. Heat map representation of DEGs (741) in WGCNA module2 in which the Co gene is a member. 639 (86.2%) of the 741 member genes in module2 were downregulated in Std2 while 73 (9.9%) were upregulated and 29 (3.9%) were unchanged (fold change <1.50). The hierarchical cluster trees indicate the relationships among genes (left) and samples (top), respectively. The broken-line boxes show the section containing the *Co* gene (indicated by a red arrow) in the overall and zoom-in views, respectively.

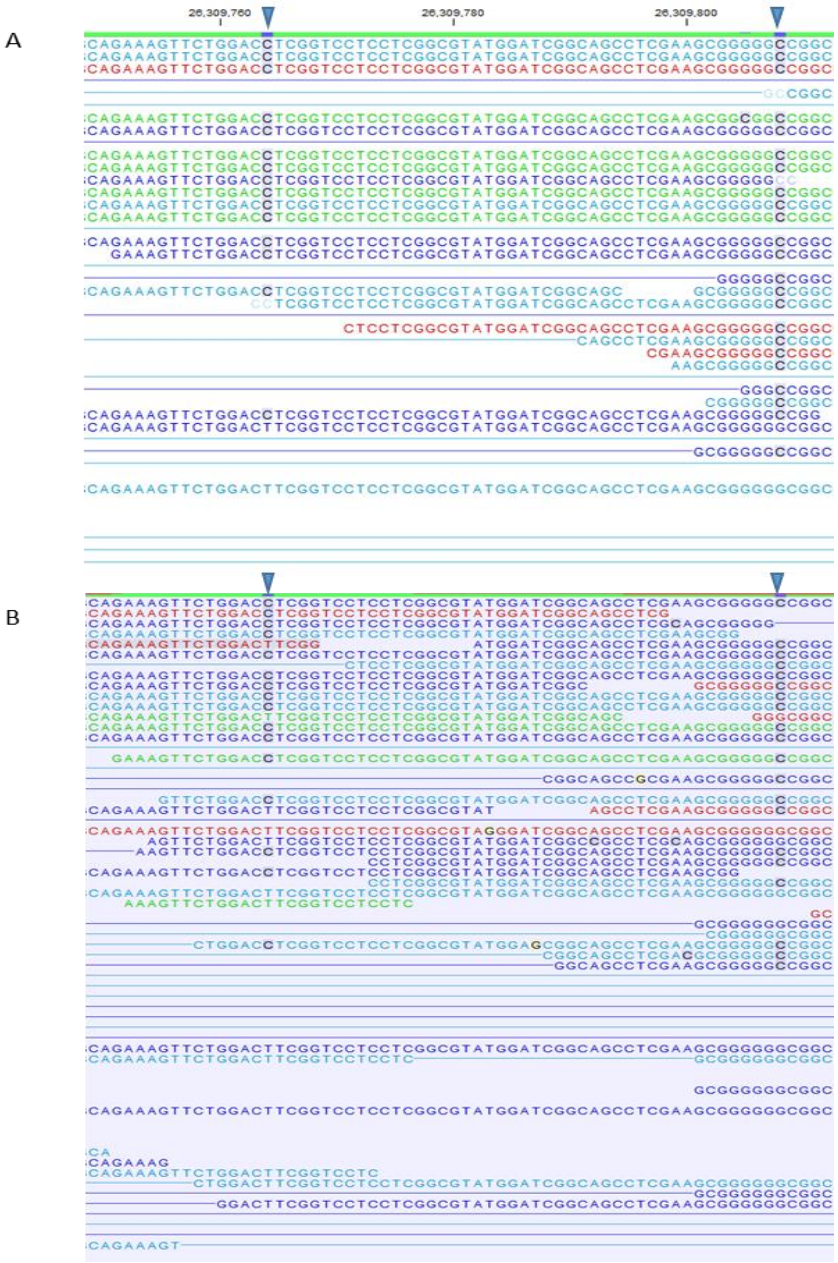


Figure S4.8. A screen snapshot of reads mapping in pools Std2 (**A**) and columnar (**B**) in a coding region of gene MD10G1170600 (AU223548) at locus *c2*. Note that the two SNVs (indicated by arrows) are homozygotes in pool Std (**A**) while heterozygous in pool columnar (**B**).

Figure S4.9. A screen snapshot of reads mapping in pools Std2 (**A**) and columnar (**B**) in a coding region of gene MD09G1171900 at locus *c3*. Note that the two SNVs and two non-SNVs (indicated by arrows) are homozygotes in pool Std (**A**) while heterozygous in pool columnar (**B**).

Supplementary Tables

Table S4.1. F_1 progeny used in pooled genome sequencing and their genotypes at the *Co* locus and growth habit.

Progeny #	Co Genotype	Growth habit-2009	Growth habit-2011	Growth habit-2015	Pool
2212	**CoCo**	C	C	C	Columnar
2237	**CoCo**	C	C	C	Columnar
2265	Coco	C	C	C	Columnar
2241	Coco	C	C	C	Columnar
2213	Coco	C	C	C	Columnar
1162	Coco	C	C	C	Columnar
2248	Coco	C	C	C	Columnar
2214	Coco	C	C	C	Columnar
2233	Coco	C	C	C	Columnar
2238	Coco	C	C	C	Columnar
2240	Coco	C	C	C	Columnar
2243	Coco	C	C	C	Columnar
2244	Coco	C	C	C	Columnar
2264	Coco	C	C	C	Columnar
2266	Coco	C	C	C	Columnar
2229	Coco	C	C	C	Columnar
1178	Coco	C	C	C	Columnar
1179	Coco	C	C	C	Columnar
2279	Coco	S	S	SL	Std2
2290	Coco	S	S	SL	Std2
2298	Coco	S	S	SL	Std2
2304	Coco	S	S	SL	Std2
2253	Coco	S	S	S	Std2
2252	Coco	S	S	S	Std2
2258	Coco	S	S	S	Std2
2269	Coco	S	S	S	Std2
2281	Coco	S	S	S	Std2
2287	Coco	S	S	S	Std2
2316	Coco	S	S	S	Std2
1159	Coco	S	S	S	Std2
1131	Coco	S	S	S	Std2
2313	Coco	S	S	S	Std2
2273	**CoCo**	S	S	S	Std2
2321	**CoCo**	S	S	S	Std2

C: columnar; S: standard; SL: standard like. Std2: Standard2

Table S4.2. Illumina raw and clean reads obtained, and stats of read mapping against the apple reference genome.

Pool		Count	Percentage of reads	Average length	Number of bases	Percentage of bases	Mean read length (bases)	seq. depth (x)
Standard2	raw reads	166,523,370			25,145,028,870		151	35.4
	clean reads	163,545,102	100.00%	124.95	20,435,067,649	100.00%	125.0	28.8
	Mapped reads	122,944,524	75.17%	129.63	15,937,142,369	77.99%		22.5
	Not mapped reads	40,600,578	24.83%	110.78	4,497,925,280	22.01%		
	Reads in pairs	90,133,668	55.11%	446.75	12,095,345,261	59.19%		
	Broken paired reads	32,810,856	20.06%	117.09	3,841,797,108	18.80%		
columnar	raw reads	259,133,758			39,129,197,458		151	55.1
	clean reads	253,038,996	100.00%	125.8	31,833,442,891	100.00%	125.8	44.9
	Mapped reads	187,827,369	74.23%	130.6	24,529,375,576	77.06%		34.6
	Not mapped reads	65,211,627	25.77%	112.01	7,304,067,315	22.94%		
	Reads in pairs	138,048,096	54.56%	402.77	18,736,289,830	58.86%		
	Broken paired reads	49,779,273	19.67%	116.38	5,793,085,746	18.20%		

Reference genome size: 709,561,391 bp

Table S4.3. Genotypes of variants common to both pools and variant segregation type inferred (with heterozygous parents)

Inheritance model	Variant genotype group	Variant genotypes observed[a]					Inferred[b]							Notes
		Std2 pool	Col pool	# of variants (1,997,962 in total)	% of variants (freq.)	Segregation types (genotype of parents)[c]	Std2 pool genotype	Std2 pool mean AF (%)	Col pool genotype			Col pool mean AF (%)	AFDD between Std2 and col pools (percentage points)	
one recessive gene	G1	He-Std2	He-Col	1,636,085	81.89	<**hh** x kk>	**h**k	50	2**h**k	**h**k		50	0	Not informative (in variants common to both pools)
						<hh x **kk**>	h**k**	50	2h**k**	h**k**		50	0	Not informative (in variants common to both pools)
						<**l**m x mm>	**l**m	50	**l**m	mm	mm	16.7	33.3	Informative for Std2 (in variants common to both pools)
						<l**m** x **mm**>	l**m**	50	l**m**	**mm**	**mm**	83.3	-33.3	Informative for Col (in variants common to both pools)
						<pp x **np**>	**n**p	50	pp	pp	**n**p	16.7	33.3	Informative for Std2 (in variants common to both pools)
						<**pp** x n**p**>	n**p**	50	**pp**	**pp**	n**p**	83.3	-33.3	Informative for Col (in variants common to both pools)
	G2	Ho-Std2	He-Col	70,522	3.53	<**h**k x **h**k>	**hh**	100	2**h**k	kk		33.3	66.7	Informative for Std2 (in variants common to both pools)
						<h**k** x h**k**>	hh	0	2h**k**	**kk**		66.7	-66.7	Informative for Col (in variants specific to pool columnar)
						<**l**m x **ll**>	**ll**	100	**ll**	m**l**	m**l**	66.7	33.3	Informative for Std2 (in variants common to both pools)
						<l**m** x ll>	ll	0	ll	**m**l	**m**l	33.3	-33.3	Informative for Col (in variants specific to pool columnar)
						<**nn** x **n**p>	**nn**	100	**nn**	**n**p	**n**p	66.7	33.3	Informative for Std2 (in variants common to both pools)
						<nn x n**p**>	nn	0	nn	np	n**p**	33.3	-33.3	Informative for Col (in variants specific to pool columnar)
	G3[e]	He-Std2	Ho-Col	39,075	1.96	NA								
	G4	Ho-Std2	Ho-Col	252,280	12.63	<**qq** x **qq**>	**qq**	100	**qq**	**qq**	**qq**	100	0	Not informative (in variants common to both pools)

Two recessive genes	G1	He-Std2	He-Col	1,636,085	81.89	<**hh** x kk>	**h**k	50	8**h**k	4**h**k	3**h**k	50	0	Not informative (in variants common to both pools)
						<hh x **kk**>	h**k**	50	7h**k**	4h**k**	4h**k**	50	0	Not informative (in variants common to both pools)
						<**l**m x mm>	**l**m	50	7**l**m	4m m	4mm	23.3	26.7	Informative for Std2 (in variants common to both pools)
						<l**m** x **mm**>	l**m**	50	7l**m**	4**m m**	4**mm**	76.7	-26.7	Informative for Col (in variants common to both pools)
						<pp x **n**p>	**n**p	50	8pp	4**n**p	3**n**p	23.3	26.7	Informative for Std2 (in variants common to both pools)
						<**pp** x n**p**>	n**p**	50	8**pp**	4n**p**	3n**p**	76.7	-26.7	Informative for Col (in variants common to both pools)
	G2	Ho-Std2	He-Col	70,522	3.53	<**h**k x **h**k>	**hh**	100	8**h**k	4kk	3**hh**	46.7	53.3	Informative for Std2 (in variants common to both pools)
						<h**k** x h**k**>	hh	0	8h**k**	4**kk**	3hh	53.3	-53.3	Informative for Col (in variants specific to pool columnar)
						<**l**m x **ll**>	**ll**	100	7**ll**	4m**l**	4m**l**	73.3	26.7	Informative for Std2 (in variants common to both pools)
						<l**m** x ll>	ll	0	7ll	4**m**l	4**m**l	26.7	-26.7	Informative for Col (in variants specific to pool columnar)
						<**nn** x **n**p>	**nn**	100	7**nn**	4**n**p	4**n**p	73.3	26.7	Informative for Std2 (in variants common to both pools)
						<nn x n**p**>	nn	0	7nn	4n**p**	4n**p**	26.7	-26.7	Informative for Col (in variants specific to pool columnar)
	G3[d]	He-Std2	Ho-Col	39,075	1.96	NA								
	G4	Ho-Std2	Ho-Col	252,280	12.63	<**qq** x **qq**>	**qq**	100	8**qq**	4**qq**	3**qq**	100	0	Not informative (in variants common to both pools)

[a] Homozygous (Ho): variant allele frequency (AF)>85%; Heterozygous (He): 85%> AF>15%
[b] for variants in the genomic regions responsible for phenotype standard2.
[c] The alleles in each first and third positions are designated to link to the recessive Std2 alleles (repressors of columnar) in the seed and pollen parents, respectively, and those in bold are a polymorphic variant in relation to the apple reference genome. Complex segregation types <ab x cd> and <ee x fg> involving semitaneously three or four DNA bases are not considered due to their relative low freqency in the genome. Based on allele frequency directional difference (AFDD) inferred, five segregation types <hk x hk> (A), <lm x ll> (B), <nn x np> (C), <lm x mm> (D), and <pp x np> (E) were considered informative for mapping the recessive traits in apple under the model of one- or two-recessive genes. Filtering informative variants was detailed in Table S4.
[d] The existance of such vairant genotype group was considered unlikely. The variants observed were likely due to the leak-throughs of other segregation types, such as <lm x mm> and <pp x np> in the variant genotype group He-Std2/He-Col, which are expected to have high variant allele frequencies 83.3% and 73.3% (close to the 85% threshold for homozygotes) under the model of one- and two-recessive genes, respectively.
Std2: standard2; Col: columnar; AFDD: allele frequency directional difference.

Table S4.4. Filters used for identification of informative variants (two recessive genes)

Segregation types	Symbol	No. of recessive genes (model)	Std2 pool mean AF (%)			Col pool mean AF (%)	AFDD between Std2 and columna pools (pecentage points)			No. of variants indentified
			Expected	Targeted	Used	Expected	Expected	Targeted	Used	
A	<**h**k x **h**k>	1	100	≥85	≥85	33.3	66.7	56.7 to 76.7	≥43.3	7,642
		2	100	≥85		46.7	53.3	43.3 to 63.3		
B	<**l**m x **ll**>	1	100	≥85	≥85	66.7	33.3	23.3 to 43.3	16.7 to 43.3	40,166
		2	100	≥85		73.3	26.7	16.7 to 36.7		
C	<**nn** x **n**p>	1	100	≥85		66.7	33.3	23.3 to 43.3		
		2	100	≥85		73.3	26.7	16.7 to 36.7		
D	<**l**m x mm>	1	50	35 to 65	35 to 65	16.7	33.3	23.3 to 43.3	16.7 to 43.3	70,230
		2	50	35 to 65		23.3	26.7	16.7 to 36.7		
E	<pp x **n**p>	1	50	35 to 65		16.7	33.3	23.3 to 43.3		
		2	50	35 to 65		23.3	26.7	16.7 to 36.7		
Sum										118,038

For identification of DNA variants under segregation type A, the cut-off is AFDD ≥43.3, ten percentage points lower than AFDD 53.3 to accommodate variations, which is estimated under model of two recessive genes. Consequently, 7,642 informative SNVs were identified among the 70,522 variants in genotype group Ho-Std2/He-Col (G2, Table S3). For DNA variants under segregation types B-C, the cut-off is AFDD 16.7, ten percentage points lower than AFDD 26.7 to accommodate variations, which is estimated under model of two recessive genes. This led to identification of 40,166 SNVs in variant genotype group Ho-Std2/He-Col (G2, Table S3). For segregation types D-E, 70,230 of the 1,636,085 SNVs in variant genotype group He-Std2/He-Col (G1, Table S3) were identified. These SNVs were obtained using the following filters: 1) the variant AF range is from 35% to 65% in pool standard2, close to their estimated mean 50%. 2) The AF is no higher than 33.3% in pool columnar, ten percentage points higher than 23.3% estimated for two recessive genes. 3) The cut-off is AFDDis 16.7 between pools standard2 and columnar.

Table S4.5. List of primers

Name of markers or target genes	Forward Primer (5' to 3')	Reverse Primer (5' to 3')	Genome location	Purpose
CH02c11	TGAAGGCAATCACTCTGTGC	TTCCGAGAATCCTCTTCGAC	Ch10: 24,261 kb	Confirmation of c2-SSR
Ch10_24818	ACCAAACCAAGACACATGCT	GGGGTTATTTACTGTGGTGGTG	Ch10: 24,818 kb	Confirmation of c2-SSR
AU223548SSR	ACCACCACTGCAGAGACTCA	GACGCACCCATTCATCTTTT	Ch10: 26,353 kb	Confirmation of c2-SSR
CH05c07	TGATGCATTAGGGCTTGTACTT	GGGATGCATTGCTAAATAGGAT	Ch09: 12,363 kb	Confirmation of c3-SSR
Hi05e07	CCCAAGTCCCTATCCCTCTC	GTTTATGGTGATGGTGTGAACGTG	Ch09: 14,303 kb	Confirmation of c3-SSR
13C2_30348-HRM	TACTTTAGCACCACCTTGTT	TGCCCGTTTAGTATATCACC	Ch09: 15,681,366	Confirmation of c3-HRM
C4935	TTTCCCAGCTGAAAAACTCG	GCAGAGAAATCCGCAGAAAC	Ch09: 17,787 kb	Confirmation of c3-SSR
CH04c07	GGCCTTCCATGTCTCAGAAG	CCTCATGCCCTCCACTAACA	Ch14: 24,205 kb	Confirmation of c4-SSR
C1374	CGGATCACAGACGCCAT	GCGTCATTTCAACAGCTTCA	Ch14: 24,421 kb	Confirmation of c4-SSR
C14087	CACCGCGTCAAAAATACCTT	CTTGTTGTTTCCCTCCCAAA	Ch06: 4,908 kb	Confirmation of c5-SSR
Hi08g03	ATTCACTTCCACCGCCATAG	GTTTGGAATGATTGCGAGTGAAGC	ch06: 6,119 kb	Confirmation of c5-SSR
CH03d07	CAAATCAATGCAAAACTGTCA	GGCTTCTGGCCATGATTTTA	ch06: 8,113 kb	Confirmation of c5-SSR

CH01h10	TGCAAAGATAGGTAGATATATGCCA	AGGAGGGATTGTTTGTGCAC	Ch08: 27,444 kb	Confirmation of c6-SSR
C13470	TCGATTCCTCAATCTCTCTCA	ATCGGAGAAAACCCAAATCC	Ch08: 30,478 kb	Confirmation of c6-SSR
MD10G1185400-Co	ATGGAGACATTAGATCAGAATCTTGT	CCATGATTGAAGACCTGGAAAAATCCG		qRT-PCR
MD17G1272100	GAGCCATCTTCCTGGGATT	CCCACCATGCATTCACTTT		qRT-PCR
MD07G1306900	TAAATGTGGAGGGAGGAGTTTT	TCTGAATTTCCTCCCACTTTCT		qRT-PCR
MD04G1096200	AGCGATTTTCGCTGAAGTG	TCAATCTGTCCAGGGTGGT		qRT-PCR
MD08G1025800	AGCACCTGGACGATCTGAC	TGCTGGGTGGTGATGTTTAT		qRT-PCR
MD01G1001600-Actin	GGCTGGATTTGCTGGTGATG	TGCTCACTATGCCGTGCTCA		qRT-PCR

Table S4.6. RNA-seq samples and statistics

Progeny #/Sample name	Co genotype	Phenotype	Mapped reads		unique reads		non-specifically		Unmapped reads		Total reads		Raw reads	Removed raw reads
			Count	%	Count	%	Count	%	Count	%	Count	%	Count	%
2230	*Coco*	Columnar	36,969,406	83.58	35,412,001	80.05	1,557,405	3.52	7,265,384	16.42	44,234,790	100	49,654,918	10.92
3213	*Coco*	Columnar	6,022,338	83.25	5,763,783	79.68	258,555	3.57	1,211,721	16.75	7,234,059	100	18,141,844	60.13
3214	*Coco*	Columnar	40,154,110	81.16	38,462,890	77.74	1,691,220	3.42	9,320,090	18.84	49,474,200	100	54,239,730	8.79
3233	*Coco*	Columnar	17,009,862	83.31	16,300,008	79.83	709,854	3.48	3,408,462	16.69	20,418,324	100	27,530,597	25.83
2220	*Coco*	Standard2	17,798,795	82.98	17,080,776	79.63	718,019	3.35	3,651,246	17.02	21,450,041	100	29,454,486	27.18
2223	*Coco*	Standard2	27,163,039	83.19	26,037,801	79.74	1,125,238	3.45	5,488,445	16.81	32,651,484	100	38,412,645	15.00
3253	*Coco*	Standard2	27,403,712	83.9	26,243,485	80.35	1,160,227	3.55	5,259,717	16.1	32,663,429	100	37,466,963	12.82
3258	*Coco*	Standard2	26,822,239	83.65	25,731,388	80.24	1,090,851	3.4	5,244,456	16.35	32,066,695	100	34,627,016	7.39
3221	*coco*	Standard1	23,042,707	85.08	22,075,512	81.51	967,195	3.57	4,041,226	14.92	27,083,933	100	33,713,293	19.66
3222	*coco*	Standard1	25,921,658	85.76	24,797,687	82.04	1,123,971	3.72	4,305,075	14.24	30,226,733	100	32,925,469	8.20
3227	*coco*	Standard1	12,731,457	85.05	12,207,680	81.55	523,777	3.5	2,237,741	14.95	14,969,198	100	21,123,725	29.14
3239	*coco*	Standard1	18,276,683	84.01	17,503,737	80.45	772,946	3.55	3,479,376	15.99	21,756,059	100	32,657,561	33.38
Sum			279,316,006		267,616,748		11,699,258		54,912,939		334,228,945		409,948,247	
Mean			23,276,334	83.7	22,301,396	80.2	974,938	3.5	4,576,078	16.3	27,852,412	100.0	34,162,354	21.5
SD			9,663,852	1.2	9,255,440	1.1	408,855	0.1	2,175,319	1.2	11,795,856		10,309,416	15.1

Table S4.7. Differentially expressed genes (DEGs) among the three phenotype groups columnar, standard1 (Std1) and standard2 (Std2)

Table is provided as a supplementary spreadsheet.

Table S4.8. Genes expressed in under c2 and c3

Table is provided as a supplementary spreadsheet.

Table S4.9. Differences and similarities in informative segregation types inferred for dominant and recessive traits

	Segregation types	Symbol	Source pools	Usefulness	Comments	Reference
Recessive	A	<hk x hk>	Common to both pools	Yes	Commonly used	This study
	B	<nn x np>	Common to both pools	Yes	Hidden	
	C	<lm x ll>	Common to both pools	Yes	Hidden	
	D	<pp x np>	Common to both pools	?	Hidden	
	E	<lm x mm>	Common to both pools	?	Hidden	
Dominant	I	<lm×mm>	Dominant trait specific pool	Yes	Commonly used	Dougherty et al 2018
	II	<lm x ll>	Common to both pools	Yes	Hidden	
	III	<hk x hk>	Common to both pools	Yes	Hidden	

CHAPTER 5

Assessing the allelotypic effect of two aminocyclopropane carboxylic acid synthase encoding genes *MdACS1* and *MdACS3a* on fruit ethylene production and softening in *Malus*.

Abstract

Phytohormone ethylene largely determines apple fruit shelf life and storability. Previous studies demonstrated that *MdACS1* and *MdACS3a*, which encode 1-aminocyclopropane-1-carboxylic acid (ACC) synthases (ACS), are crucial in apple fruit ethylene production. *MdACS1* is well-known to be intimately involved in the climacteric ethylene burst in fruit ripening while *MdACS3a* has been regarded a main regulator for ethylene production transition from system 1 (during fruit development) to system 2 (during fruit ripening). However, *MdACS3a* was also shown to have limited roles in initiating the ripening process lately. To better assess their roles, fruit ethylene production and softening were evaluated at five time-points during a 20-d postharvest period in 97 *Malus* accessions and in 34 progeny from two controlled crosses. Allelotyping was accomplished using an existing marker (ACS1) for *MdACS1* and two markers ($CAPS_{866}$ and $CAPS_{870}$) developed here to specifically detect the two null alleles (*ACS3a-G289V* and *Mdacs3a*) of *MdACS3a*. In total, 952 *Malus* accessions were allelotyped with the three markers. The major findings included: The effect of *MdACS1* was significant on fruit ethylene production and softening while that of *MdACS3a* was less detectable; allele *MdACS1-2* was significantly associated with low ethylene and slow softening; under the same background of the *MdACS1* allelotypes, null allele *Mdacs3a* (not *ACS3a-G289V*) could confer a significant delay of ethylene peak; alleles *MdACS1-2* and *Mdacs3a* (excluding *ACS3a-G289V*) were highly enriched in *M. domestica* and *M. hybrid* when compared with those in *M. sieversii*. These findings are of

practical implications on utilizing the beneficial alleles *MdACS1-2* and *Mdacs3a*.

Introduction

To make fresh apple fruit available year-round for consumers, the controlled atmosphere (CA) storage technology has been adapted widely in the apple industry. The technology primarily employs low temperature, low O_2, and high CO_2 in combination with an ethylene production inhibitor 1-methylcyclopropene (1-MCP) and others. Apple fruit can be stored for over 10 months under optimal CA conditions. However, physiological disorders associated with CA storage, such as injuries induced by cold and CO_2 and flesh browning induced by 1-MCP, can cause substantial loss for storage operators (Burmeister & Dilley, 1995; Watkins et al., 2005; Watkins, Silsby, & Goffinet, 1997). Such storage disorders have been reported for major apple varieties such as 'Empire' and 'McIntosh' (Fawbush, Nock, & Watkins, 2008; Watkins et al., 1997) and for rising cultivars such as 'Honeycrisp' (Watkins et al., 2005). There is a strong need for new apples of long shelf life and improved keeping quality with few or no storage disorders.

The gaseous phytohormone ethylene pays an important role in climacteric fruit ripening. The shelf life and storability of apple fruit are closely correlated with their ethylene production levels. Plant ethylene biosynthesis has been well defined in the Yang cycle that involves three enzymes: S-adenosylmethionine (SAM) synthase, 1-aminocyclopropane-1-carboxylic acid (ACC) synthase (ACS) and ACC oxidase (ACO) (Yang & Hoffman, 1984). The enzymes ACS and ACO have been the subject of extensive studies to better understand plant ethylene production. Studies in many plant species including tomato and apple have shown that ACS and ACO are encoded by gene families of multiple members, i.e. the ACS family and the ACO family, respectively.

There are two systems of ethylene production in plants: system 1 occurs during plant/fruit growth and development, and system 2 is defined exclusively for the floral senescence and fruit

ripening stages (Barry & Giovannoni, 2007). In tomato, system 1 ethylene biosynthesis involves LeACS6, 1A and LeACO1, 3, 4; whereas system 2 uses LeACS2, 4 and ACO1, 4 (Cara & Giovannoni, 2008). In apple, at least five *ACS* (*MdACS1-5*) and four *ACO* (*MdACO1-4*) genes have been reported (Satoru Kondo, Meemak, Ban, Moriguchi, & Harada, 2009; Wiersma, Zhang, Lu, Quail, & Toivonen, 2007) and these genes appear to be operating similarly in the two systems for ethylene production. Figure 5.1 shows the inferred pathway in apple, based on multiple studies.

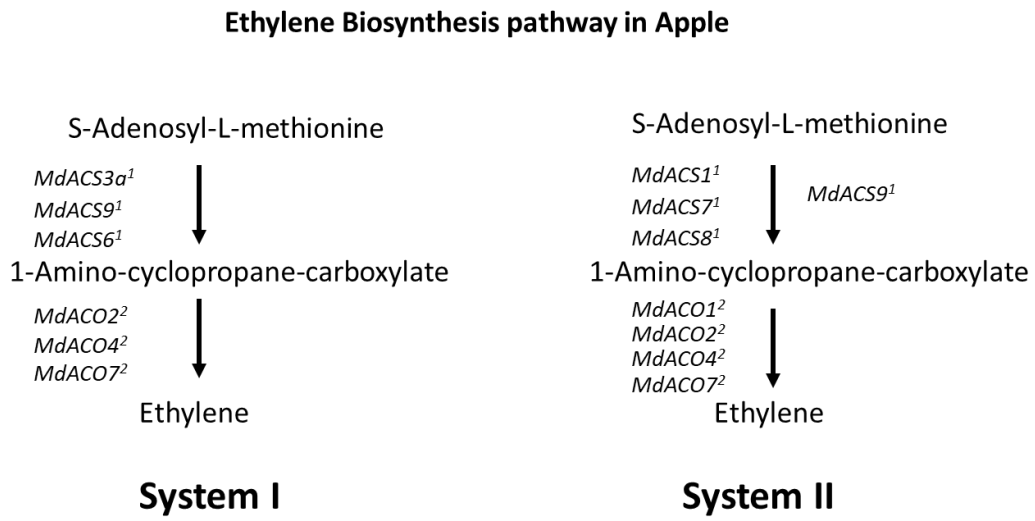


Figure 5.1. Inferred ethylene biosynthesis pathway in apple after S-Adenosyl-L-methionine synthesis. System I and System II are shown. Different MdACS (1-amino-cyclopropane-carboylate-synthase) and MdACO (1-amino-cyclepropane-carboylate-oxidase) genes known to be involved are shown. 1 (Li, Tan, Yang, & Wang, 2013) 2 (Singh, Weksler, & Friedman, 2017).

MdACS1 is considered a system 2 gene; and its expression is highly correlated with the ethylene production burst in ripening apples. There are two alleles for the *MdACS1* gene, *MdACS1-1* and *MdACS1-2*, and the former is often associated with high ethylene production while the latter with lower ethylene during fruit ripening (Costa et al., 2005; Harada et al., 2000; Oraguzie, Iwanami, Soejima, Harada, & Hall, 2004; Sato et al., 2004; Sunako et al., 1999). This

observation has led to a marker assisted selection strategy emphasizing on selection for allelotype (see Discussion for usage of term 'allelotype') *MdACS1-2/2* for long shelf life apples (Zhu & Barritt, 2008). Indeed, some evidence suggests that modern apple breeding practice has unintentionally favored selection for the *MdACS1-2* allele in commercial apple cultivars (H. Nybom, Sehic, & Garkava-Gustavsson, 2008), presumably for fruit of low ethylene and long shelf life.

However, early-ripening cultivars showed faster fruit softening, regardless of their *MdACS1* allelotypes (Harada et al., 2000). This is consistent with the observation that the polygalacturonase gene (*MdPG1*) involved in softening c v. Therefore, there are other factors also affecting fruit shelf life in addition to *MdACS1*. Interestingly, findings in a recent report have suggested that allele variations of another *ACS* gene (U73816) (Rosenfield, Kiss, & Hrazdina, 1996), designated *MdACS3a* (AB243060), are an essential factor regulating apple fruit ripening and shelf life (Aide Wang et al., 2009). There are two natural mutant alleles of the wild type allele *MdACS3a*: One is the functional null allele *MdACS3a-G289V*, arising from a point mutation that leads to an amino acid substitution from G_{289} to V_{289} at an active region for the MdACS3A enzyme activity, resulting in a functionally inactive enzyme. In melon, a similar point mutation in a conserved active region of an *ACS* gene led to andromonoecy, a common sexual system in angiosperms by plants characterized carrying both male and bisexual flowers (Boualem et al., 2008). This is an excellent example demonstrating that point mutations in conserved active regions of an ACS enzyme could confer a major phenotypic variation in plants. The other, a transcriptionally null allele *Mdacs3a*, is characterized by non-detectable mRNA. Moreover, combinations of *Mdacs3a* and *MdACS3a-G289V* alleles, regardless homozygous or heterozygous, are highly associated with lower ethylene production and long shelf life. In the six

apple varieties/selections of the two null alleles studied, all showed low ethylene production and long shelf life, irrespective to their *MdACS1* allelotypes and early, mid or late physiological maturation dates (Aide Wang et al., 2009). Furthermore, the expression of *MdACS3a* is fruit tissue specific and detectable only during the transition from system 1 to system 2 ethylene biosynthesis (Satoru Kondo et al., 2009; Aide Wang et al., 2009; Wiersma et al., 2007). These observations suggest that *MdACS3a* acts like a main regulator for the transition, thereby crucial in regulating the fruit ripening process (Aide Wang et al., 2009).

In a more recent report, however, the allelotypes of *MdACS3a* were demonstrated to affect the ripening initiation of late maturing cultivars only, but not the early- or mid-maturing cultivars (S. Bai et al., 2012). To better assess the roles of *MdACS1* and *MdACS3a*, two approaches were taken in this study. The first approach was to estimate the allelotypic effect of the two genes by evaluating fruit ethylene production levels and softening rates in 97 diverse *Malus* accessions and 34 progeny from two controlled crosses. The second approach was to examine how variations in their allelotypic effect were associated with the frequency changes of the *MdACS1* and *MdACS3a* alleles in *M. domestica* and *M. hybrid* as compared with those in *M. sieversii*, the major progenitor species of domestic apples, in 952 *Malus* accessions covering 53 *Malus* species. Allelotyping (see Discussion for usage of term 'allelotyping') of *MdACS1* and *MdACS3a* was conducted using an existing marker for *MdACS1* and two CAPS (cleaved amplified polymorphic sequence) markers specifically developed here to detect alleles *ACS3a-G289V* and *Mdacs3a*.

Materials and Methods

Plant materials

Two sets of *Malus* accessions were used in this study, which have been planted and maintained in the *Malus* germplasm repository of the U.S. Department of Agriculture (USDA) in Geneva, New York. The first set included a total of 952 accessions, covering 53 *Malus* species (**Table S5.1**). Among them, *Malus domestica* of 508 accessions, *M. hybrid* (the breeding selections derived from crosses between *M. domestica* and other *Malus* species) of 146, and *M. sieversii* (the major progenitor species of *M. domestica*) of 78 were most commonly represented (**Table S5.1**). The second set comprised 34 half-sib progeny selected from two interspecific crosses GMAL4592 ('Royal Gala' × PI613978) and GMAL4593 ('Royal Gala' × PI613981). 'Royal Gala', a widely grown apple cultivar (*M. domestica*), has an allelotype *MdACS1-2/2* and *MdACS3a/MdACS3a-G289V* for genes *MdACS1* and *MdACS3a*, respectively. PI613978 and PI613981 are among the elite selections of *M. sieversii* collected from Kazakhstan (Forsline, Aldwinckle, Dickson, Luby, & Hokanson, 2003), and they have the same allelotypes for the two ACS genes, i.e. *MdACS1-1/1* and *MdACS3a/MdACS3a-G289V*. Population GMAL4592 was used in one of our previous studies (Y. Bai et al., 2012). Both GMAL4592 and GMAL4593 were planted on their own seedling roots in 2004.

Measurements of fruit ethylene production and firmness

Fruit ethylene production and flesh firmness were measured for 97 of 952 *Malus* accessions in the first set and the 34 half-sib progeny in the second set as described previously (A. Wang & Xu, 2012). Briefly, for each accession, at least 25 fruit were harvested at a target maturity level as determined by the starch index of 4-6 according to the Cornell Starch Chart

(Blanpied & Silsby, 1992). The 25 fruit were evenly divided into five groups and were stored for 0, 5, 10, 15 and 20 days at room temperature (20-25ºC), respectively. Each fruit was weighed then enclosed in a gas-tight container (1.2 L) and kept for 1 h at room temperature. One milliliter (mL) of gas was sampled from the headspace in the container using a BD syringe (No. 309602, Franklin Lakes, NJ). The gas sample's ethylene concentration was measured with a gas chromatograph (GC) HP 5890 series II (Hewlett-Packard, Palo Alto, CA) equipped with a flame ionization detector. Before the gas samples were assayed, the GC was calibrated with a standard ethylene gas (NO. 34489, Restek, Bellefonte, PA) at a series of concentrations 0.01, 0.1, 0.5, 1, 5, 10, and 100 ppm, respectively, to obtain the linear relation between ethylene peak area and concentration. The fruit ethylene production was calculated with the following formula:

$$E=[C_2H_4]\times(V_1-V_2)/W/T$$

Where E stands for fruit ethylene production rate in nanoliter (nL) per gram (g) of fresh weight per hour (h), $[C_2H_4]$ for ethylene concentration in ppm, V_1 for the volume of container in mL, V_2 for the volume of fruit in mL equivalent to fresh weight (W) in grams, and T stands for the time in hours kept in the container.

Fruit flesh firmness was measured using a penetrometer (Fruit Tester, Wagner FTK100, Greenwich, CT) with a probe of 11-mm in diameter. The probe tip was pressed vertically into the fruit pulp (after skin removal) to a depth of 10-mm. For larger fruit, four skin discs were removed from opposite sides of each fruit along the equator, and for smaller fruit, three skin discs were removed at roughly equal distance. The firmness readings were expressed in kg/cm^2, and firmness loss was measured by the percentage (%) of firmness reduced at d5-d20 as compared with the firmness at d0. After the firmness was measured, fruits were sliced in half along the equator, dipped into a iodine-potassium iodide (I_2-KI) solution, and then allowed the

staining reaction for >1 min before reading Cornell Starch Index (Blanpied & Silsby, 1992).

Allelotyping of *MdACS1* and *MdACS3a*

Allelotyping of *MdACS1* was conducted with marker ACS1 using primers ACS1-5F/R (Table S2) as reported previously (Harada et al., 2000; Zhu & Barritt, 2008). However, Allelotyping of *MdACS3a* was accomplished with two cleaved amplified polymorphic sequences (CAPS) markers developed in this study using an online tool for identifying appropriate restriction enzymes (Neff, Turk, & Kalishman, 2002) (see Results). These two markers, named $CAPS_{866}$ and $CAPS_{870}$, were capable of detecting the functional null allele *MdACS3a-G289V* and the transcriptional null allele *Mdacs3a*, respectively. In practice, the same primers ACS3a-289F/R (**Table S5.2**) were used for polymer chain reactions (PCR) to amplify the targeted DNA fragment for both $CAPS_{866}$ and $CAPS_{870}$. PCRs were performed with 35 cycles of 94 °C for 30 s, 58 °C for 30 s, 72 °C for 1 min, with an initial 94 °C for 5 min and a final extension of 72 °C for 10 min. Each PCR reaction mix was set in 10 μl containing 20 ng genomic DNA, 0.2 mM each dNTP, 0.5 μM of each primer, 2.5 mM $MgCl_2$, 2 μl 5× PCR Colorless GoTaq Reaction Buffer and 1 U of GoTaq DNA polymerase (Promega, Madison, WI). To detect alleles *MdACS3a-G289V* and *Mdacs3a*, the PCR products were restricted with enzymes BstNI and Taq$^{\alpha}$I (New England Biolabs, Ipswich, MA) following the manufacturer's instruction, respectively. The restricted PCR products were assayed by electrophoresis on 1.5% agarose gel and then stained with ethidium bromide for visualization and documentation as described previously (Y. Bai et al., 2012).

Sanger DNA sequencing

The PCR products amplified by primers ACS3a-289F/R (Table S2) were directly sequenced using a DNA sequencer ABI3730XL (Applied Biosystems, Foster City, CA) at the Cornell University Biotechnology Resource Center (Ithaca, NY). The reverse PCR primer ACS3a-289R was used for DNA sequencing. DNA sequence analyses were performed using software Sequencher 5.2 (Gene Codes Corporation, Ann Arbor, MI).

Statistical analysis

Pearson's correlation analysis and one-way analysis of variance (ANOVA) of ethylene production and fruit firmness were conducted with software JMP® Pro 10.0 (SAS institute Inc., Cary, NC). Significance levels in comparison of the means were determined by $p<0.05$ (Student's T test).

Results

Evaluation of fruit ethylene production and softening

Fruit ethylene production and softening were evaluated in 97 of 952 *Malus* accessions (**Table S5.1**, **Table S5.3**). Their maturity date was determined by Cornell starch index, which had a mean 5.5±1.4 at harvest. The 97 accessions varied widely not only in maturity date (from August 16th through November 8th, 2011) (**Figure S5.1A**) and fruit weight (25.1-303.8g, Fig.S1b), but also in ethylene production and firmness at harvest (d0) and during the 20-d postharvest period (**Figure 5.2 A, B**). At d0, for example, the ethylene levels ranged from 0.7 nL/g/h of PI588844 ('Fuji', *M. domestica*) to 679.3 nL/g/h of PI619168 (an accession of *M. sylvestris*), and fruit firmness varied from 3.8 kg/cm^2 of PI589572 (E14-32, *M. hybrid*) to 12.7 kg/cm^2 of PI589478 ('Novosibirski Sweet', *M. domestica*). Despite being highly variable, a trend

line of bivariate function could be fit for fruit ethylene production (r^2=0.120, p<0.0001, **Figure 5.2 A**) and fruit firmness (r^2=0.147, p<0.0001, **Figure 5.2 B**).

The trend line of fruit ethylene showed a peak between d10 and d15, which was largely a reflection of the mean fruit ethylene levels 75.5±100.5, 207.3±193.9, 272.8±249.6, 247.0±170.8 and 217.3±146.5 (nL/g/h) at d0, d5, d10, d15 and d20, respectively (**Figure 5.2A**). A majority (59/97, 60.8%) of the 97 *Malus* accessions reached their peak ethylene day at d10 (25) or d15 (34) while 2, 16 and 20 accessions topped their ethylene production at d0, d5 and d20 (**Figure S5.1 C**). The peak ethylene reads were spread from 1.7 nL/g/h of PI589570 (E36-7, *M. hybrid*) at d20 to 1022.2 nL/g/h of PI633801 (*M. sieversii*) at d10 (**Table S5.3**).

As expected, the trend line of fruit firmness showed a continuous decreasing during the 20-d period (**Figure 5.2 B**). This was also an approximation of the mean firmness 7.4±1.7, 6.5±2.1, 5.8±2.0, 5.3±1.99 and 5.3±1.92 (kg/cm^2) at d0, d5, d10, d15 and d20, respectively. In other words, the mean fruit firmness was lost by 13.6% at d5, 22.0% at d10, 29.2% at d15 and 29.0% at d20.

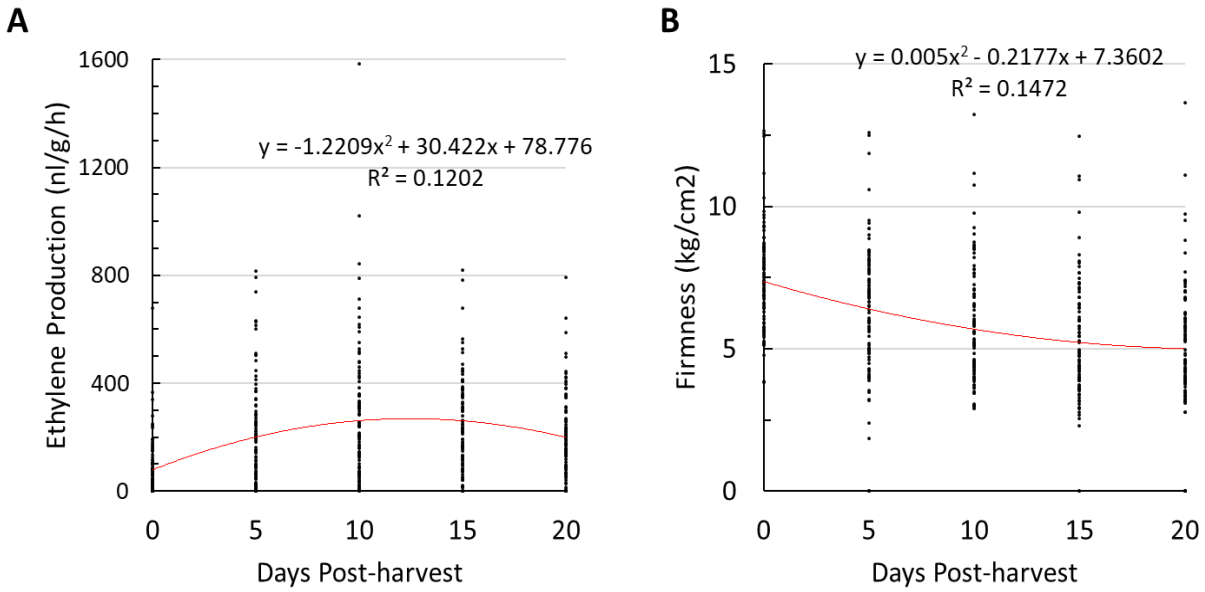


Figure 5.2. Evaluation of fruit ethylene production (**A**) and firmness (**B**) in 97 *Malus* accessions during a 20-d postharvest period under room temperature. The trend lines (curves in red) and the associated equations and coefficient of determination (R^2) are presented.

Fruit ethylene production and firmness loss was significantly correlated (**Table 5.1**). The strongest correlation (r=0.564, p=0) was observed between ethylene at d15 and fruit firmness loss at d10 while the weakest (r=0.214, p=0.035) was between ethylene at d10 and fruit firmness loss at d5. Peak ethylene day (day of peak ethylene production during the 20-d post-harvest storage) was most significantly correlated with ethylene at d5 (r=-0.479, p=6.9E-7), and it also significantly correlated with fruit firmness loss at d10 (r=-0.258, p=0.011) and d15 (r=-0.238, p=0.019) (**Table 5.1**).

Table 5.1. Correlation coefficients between fruit ethylene production and firmness or firmness loss in 97 *Malus* accessions[a].

	C_2H_4-d0	C_2H_4-d5	C_2H_4-d10	C_2H_4-d15	C_2H_4-d20	Firmness d0 (kg/cm^2)	Firmness loss_d5 (%)	Firmness loss_d10 (%)	Firmness loss_d15 (%)	Firmness loss_d20 (%)	Peak C_2H_4 day[b]
C_2H_4-d0	1.000**										
C_2H_4-d5	0.434**	1.000**									
C_2H_4-d10	0.410**	0.695**	1.000**								
C_2H_4-d15	0.353**	0.734**	0.839**	1.000**							
C_2H_4-d20	0.265**	0.733**	0.695**	0.871**	1.000**						
Firmness d0 (kg/cm^2)	-0.208*	-0.261**	-0.164	-0.239*	-0.220*	1.000**					
Firmness loss_d5 (%)	0.324**	0.484**	0.214*	0.334**	0.431**	-0.198	1.000**				
Firmness loss_d10 (%)	0.345**	0.538**	0.493**	0.564**	0.478**	-0.092	0.704**	1.000**			
Firmness loss_d15 (%)	0.284**	0.446**	0.421**	0.481**	0.442**	-0.104	0.699**	0.863**	1.000**		
Firmness loss_d20 (%)	0.306**	0.453**	0.388**	0.481**	0.438**	-0.041	0.596**	0.828**	0.853**	1.000**	
Peak C_2H_4 day[b]	-0.229*	-0.479**	-0.402**	-0.281**	-0.211*	0.219*	-0.112	-0.258*	-0.238*	-0.190	1.000**

[a] Fruit firmness loss was measured in a 20-d postharvest period under room temperature. C_2H_4: ethylene; [b] Peak C_2H_4 (ethylene) day: day of peak ethylene production during the 20-d post-harvest storage; asterisk stars * and ** stand for significance levels exceeding p0.05 (r=0.1996, n=97) and p0.01(r=0.2603, n=97), respectively.

Development of allelic specific markers for *MdACS3a*

The null allele *MdACS3a-G289V* is caused by a mutation from G_{866} to T_{866} at the 866th base in the coding sequence of *MdACS3a* (Aide Wang et al., 2009). Based on the web-based tool for single nucleotide polymorphism (SNP) analysis (Neff et al., 2002), the mutation abolishes the recognition site CC_{866}WGG of restriction enzyme *BstNI* (**Figure 5.3**). To develop a CAPS marker, two primers (ACS3a289F/R) (**Table S5.2**) were designed to amplify a DNA fragment (480 bp) covering the SNP (G_{866}/T_{866}) specifically from *MdACS3a* although the three *MdACS3* member genes *MdACS3a* (AB243060), *MdACS3b* (AB243061) and *MdACS3c* (AB243062) are

of high identity in their DNA sequences (Aide Wang et al., 2009). The specificity of the primer pair to *MdACS3a* was confirmed by sequencing of the PCR products from 92 of the 97 *Malus* accessions (**Figure 5.3**, **Table S5.3**). Digestion of the PCR products with *BstNI* yielded restriction bands as expected (**Figure 5.4 B**), indicating the successful development of a CAPS marker detecting SNP G_{866}/T_{866}, designated $CAPS_{866}$. Therefore, allele $CAPS_{866}G$ represents the wide type allele *MdACS3a* while $CAPS_{866}T$ stands for the functional null allele *MdACS3a-G289V*.

Development of a marker detecting the transcriptional null allele *Mdacs3a* was initially thought to be challenging as the null allele was reported not to show sequence variations from the wild type allele (Aide Wang et al., 2009). However, sequencing analysis of the PCR products amplified by primers ACS3a289F/R in the 92 accessions (**Table S5.3**) not only identified the expected SNP G_{866}/T_{866}, but also a new SNP C_{870}/T_{870} (Fig.2). Importantly, this new SNP discriminates the two alleles of *MdACS3a* in 'Fuji' (Fig.2), which was known of allelotype *MdACS3a*/*Mdacs3a* (Aide Wang et al., 2009). Evidence from this and other studies (see Discussion) indicated that base T_{870} was associated with the *Mdacs3a* allele. Using a similar approach, another CAPS marker, named $CAPS_{870}$ was developed to detect SNP C_{870}/T_{870} using restriction enzyme $Taq^{\alpha}I$ along with the same primers ACS3a289F/R (**Figure 5.4 C**). Therefore, allele $CAPS_{870}C$ corresponds to the wide type allele *MdACS3a* while $CAPS_{870}T$ to the transcriptional null allele *Mdacs3a*.

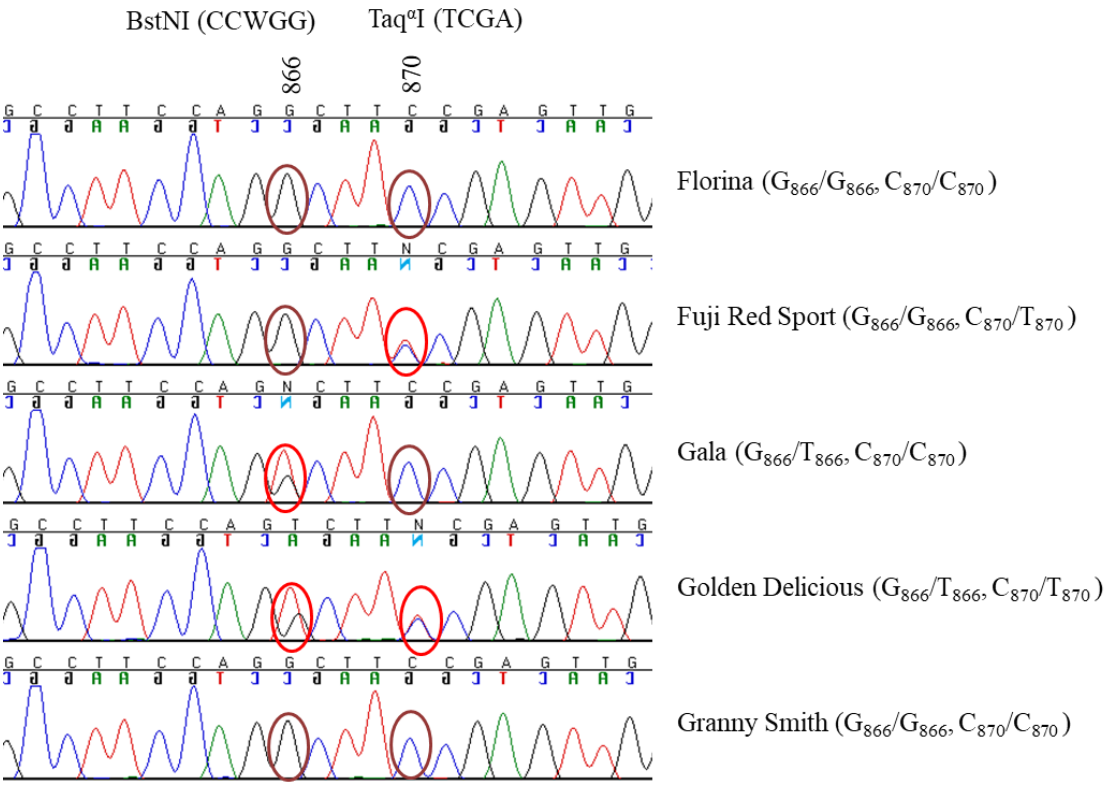


Figure 5.3. Chromatogram of the DNA sequence (partial) of *MdACS3a* encompassing SNPs G866/T866 and C870/T870 in six apple cultivars 'Florina', 'Fuji red sport', 'Gala', 'Golden Delicious' and 'Granny Smith'. The oval circles in brown and red indicate the homozygous or heterozygous status at the 866th and 870th nucleotides in the coding sequence of *MdACS3a*, respectively. The recognition sites of restriction enzymes *BstNI* and *Taq$^{\alpha}$I* are provided to show that the mutation from G_{866} to T_{866} abolishes the restriction site of *BstNI* while the mutation from C_{870} to T_{870} gives rise to a restriction site for *Taq$^{\alpha}$I*. The right panel shows allelotypes of *MdACS3a* as represented by the SNP alleles, where G_{866} stands for allele *MdACS3a* (wild type), T_{866} for *MdACS3a-G289V* (functional null allele), C_{870} also for allele *MdACS3a*, and T_{870} for *Mdacs3a* (transcriptional null allele).

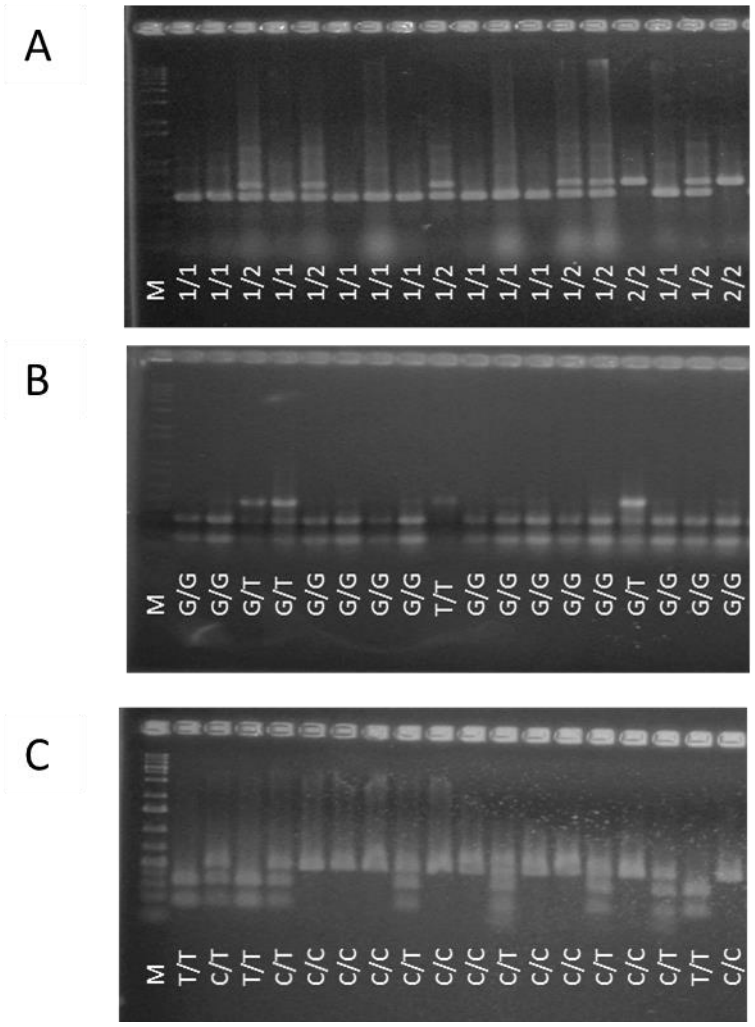


Figure 5.4. Agarose gel analyses of markers ACS1 (**A**), CAPS866 (**B**) and CAPS870 (**C**). For marker ACS1, the PCR products amplified by primers ACS1-5 F/R were directly analyzed. Allelotypes *MdACS1-1/1*, *MdACS1-2/2* and *MdACS1-1/2* are denoted with '1/1', '2/2' and '1/2', respectively. For marker $CAPS_{866}$, the PCR products were first amplified by primers ACS3a-289F/R and then digested with enzyme *BstNI*, which restricts the *MdACS3a* (G_{866}) allele into the two lower bands. Allelotypes *MdACS3a*/*MdACS3a* (G_{866}/G_{866}), *MdACS3a*/*MdACS3a-G289V* (G_{866}/T_{866}), and *MdACS3a-G289V/G289V* (T_{866}/T_{866}) are noted with 'G/G', 'G/T' and 'T/T', respectively. For marker $CAPS_{870}$, enzyme *Taq*$^{\alpha}$*I* restricts the *Mdacs3a* (T_{870}) allele into the two lower bands. Allelotypes *MdACS3a*/*MdACS3a* (C_{870}/C_{870}), *MdACS3a*/*mdacs3a* (C_{870}/T_{870}), and *mdacs3a/mdacs3a* (T_{870}/T_{870}) are noted with 'C/C', 'C/T' and 'T/T', respectively.

Effect of the allelotypes of *MdACS1* and *MdACS3a* on ethylene production and firmness loss

To evaluate the effect of the allelotypes of *MdACS1* and *MdACS3a*, the 97 *Malus* accessions were assayed with markers ACS1, $CAPS_{866}$ and $CAPS_{870}$ that can detect different alleles of *MdACS1* and *MdACS3a* (**Figure 5.4**). As a result, marker ACS1 identified 53, 36 and 8 accessions of allelotypes of *MdACS1-1*/*MdACS1-1* (*MdACS1-1/1*), *MdACS1-1*/*MdACS1-2* (*MdACS1-1/2*) and *MdACS1-2*/*MdACS1-2* (*MdACS1-2/2*), respectively (**Table S5.3**). Similarly, marker $CAPS_{866}$ detected 75 accessions of allelotype $CAPS_{866}G/CAPS_{866}G$ ($CAPS_{866}G/G$), 18 of $CAPS_{866}G/CAPS_{866}T$ ($CAPS_{866}G/T$), and four of $CAPS_{866}T/CAPS_{866}T$ ($CAPS_{866}T/T$), and marker $CAPS_{870}$ uncovered 47 accessions of allelotype $CAPS_{870}C/CAPS_{870}C$ ($CAPS_{870}C/C$), 40 of $CAPS_{870}C/CAPS_{870}T$ ($CAPS_{870}C/T$), and ten of $CAPS_{870}C/CAPS_{870}T$ ($CAPS_{870}T/T$) (**Table S5.3**).

A series of one-way ANOVA of the fruit ethylene production and fruit firmness loss over the 20-d period within each of the three allelotype groups (**Figure 5.5**) indicated that the most differences were observed among the *MdACS1* allelotypes. Allelotype *MdACS1-1/1* showed significantly higher ethylene production (d0-d20) and firmness loss (d5-d20) than *MdACS1-1/2* and *MdACS1-2/2* allelotypes, but *MdACS1-1/2* and *MdACS1-2/2* did not differ in terms of ethylene production or firmness retention (**Figure 5.5 A, D**). In contrast, there were no difference among the $CAPS_{866}$ allelotypes in fruit ethylene production and firmness loss (**Figure 5.5 B, E**). Among the $CAPS_{870}$ allelotypes, significant difference was not detected for ethylene production, but there were differences in fruit firmness loss between allelotypes $CAPS_{870}C/C$ and $CAPS_{870}C/T$ at d5 and between $CAPS_{870}C/C$ and $CAPS_{870}T/T$ at d10 (**Figure 5.5 C, F**). This indicated that such differences in fruit firmness loss at d5 and d10 in the $CAPS_{870}$ allelotypes might be caused by other factors rather than their ethylene production levels.

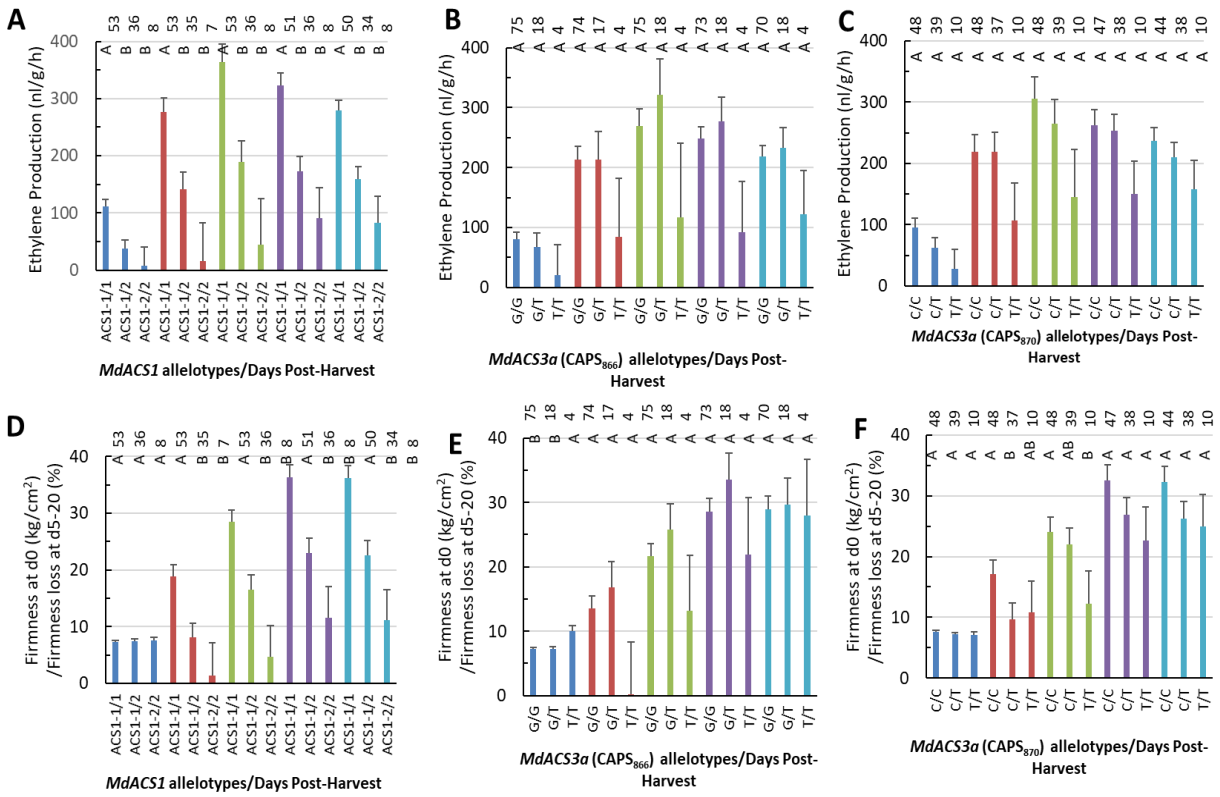


Figure 5.5. Comparison of the means of fruit ethylene production and firmness or firmness loss among allelotypes of *MdACS1* as defined by marker ACS1 (**A, D**), and among those of MdACS3a as defined by markers CAPS866 (**B, E**) and CAPS870 (**C, F**). The allelotypes are annotated similarly as those in the legend of Fig.3. Colors of column in blue, orange, green, purple and turquoise represent d0, d5, d10, d15 and d20, respectively. The statistical tests were conducted independently within each of the five storage time points (d0-d20). Significance levels are indicated with letters (shown above the columns in the chart), where different letters indicate $p<0.05$. The numbers of accessions observed (n) for each allelotype are presented accordingly (shown above the letters for significance). Error bars indicate standard errors.

To seek such factors, peak ethylene day, which measures ethylene peak timing, was examined (**Figure 5.6**) as this trait was negatively correlated with fruit firmness loss at d10 (r=-0.258, p=0.011) although the correlation was insignificant at d5 (r=-0.112, p=0.275) (**Table 5.1**). Encouragingly, the three $CAPS_{870}$ allelotypes showed significant difference from each other, with *$CAPS_{870}$C/T* having peaked the earliest, *$CAPS_{870}$C/C* intermediate, and *$CAPS_{870}$T/T* the latest (Fig.5a). These data appeared to suggest that the earlier peak ethylene day of *$CAPS_{870}$C/C* might

have contributed to its greater fruit firmness loss of *CAPS$_{870}$C/C* as compared with that of *CAPS$_{870}$T/T* at d10 (**Figure 5.5 F**). However, the lowest fruit firmness loss of *CAPS$_{870}$C/T* at d5 remained to be explained. Peak ethylene day was also analyzed in the other two groups of allelotypes. In the allelotypes of *MdACS1*, *MdACS1-1/1* had an earlier peak ethylene than *MdACS1-2/2*, but showed no difference from *MdACS1-1/2* (**Figure 5.6 A**). In the three allelotypes of CAPS$_{866}$, no significant difference was observed **(Figure 5.6 A**).

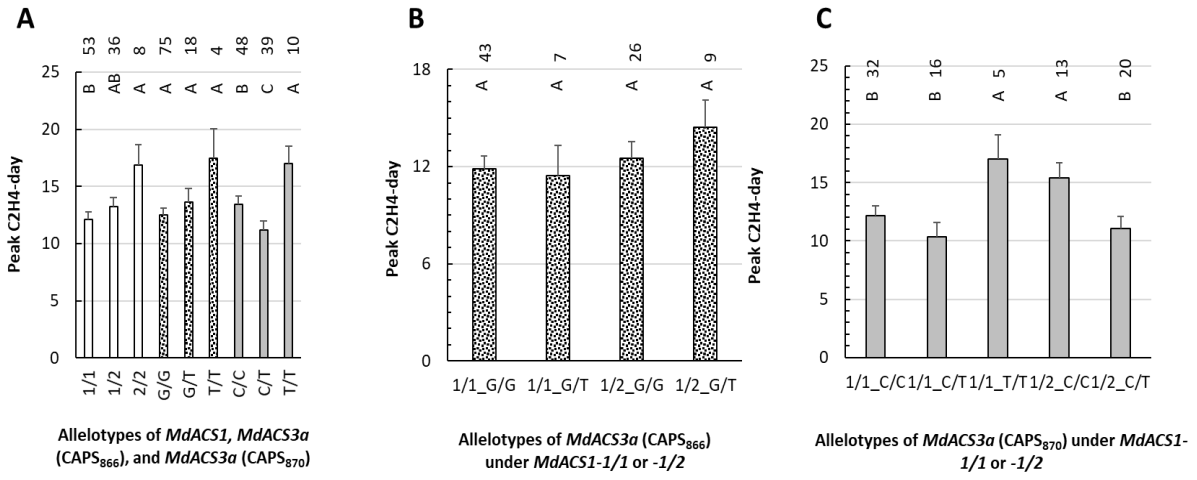


Figure 5.6. Comparison of the means of peak ethylene day among the allelotypes of *MdACS1* as defined by marker ACS1 (open column) and those of *MdACS3a* as defined by markers CAPS866 (dot-filled column) and CAPS870 (filled column) (**A**), and among the allelotypes of *MdACS3a* defined by markers CAPS$_{866}$ (**B**) and CAPS$_{870}$ (**C**) under the same background of *MdACS1-1/1* or *MdaCS1-1/2*. The allelotypes, significance levels and observed numbers are represented similarly as those in Figure 5.3 and Figure 5.4.

It was clear that the effect of *MdACS1* on ethylene production and fruit firmness loss was much stronger than that of *MdACS3a* (**Figure 5.5**). To see if the random presence of the *MdACS1* alleles might have obscured the detection of the effect of *MdACS3a* allelotypes (**Figure 5.5 B, C, E, F**), another series of ANOVA was conducted for the *MdACS3a* allelotypes of five or more accessions (**Figure 5.7**) under the same background of *MdACS1* allelotypes *MdACS1-1/1* and *MdACS1-1/2*, which occurred in 53 and 36 of the 97 accessions (**Table S5.3**), respectively. The third allelotype *MdACS1-2/2* was not included in the analysis (Fig.6) due to limited number

of accessions (8).

For $CAPS_{866}$, the ANOVA analyses were conducted for two allelotypes *$CAPS_{866}$G/G* and *$CAPS_{866}$G/T* under *MdACS1-1/1* as well as under *MdACS1-1/2* (**Figure 5.6 B**, **Figure 5.7 A, C**). This allowed us to identify that allelotype *$CAPS_{866}$G/T* produced significantly higher levels of ethylene than *$CAPS_{866}$G/G* at d10 under *MdACS1-1/1* (**Figure 5.7 A**). For $CAPS_{870}$, three allelotypes *$CAPS_{870}$C/C*, *$CAPS_{870}$C/T* and *$CAPS_{870}$T/T* under *MdACS1-1/1* and two allelotypes *$CAPS_{870}$C/C* and *$CAPS_{870}$C/T* under *MdACS1-1/2* were analyzed (**Figure 5.6 C**, **Figure 5.7 B, D**). The results showed that allelotype *$CAPS_{870}$T/T* had significant later peak ethylene day than *$CAPS_{870}$C/C* and *$CAPS_{870}$C/T* under *MdACS1-1/1*, and *$CAPS_{870}$C/C* had significant later peak ethylene than *$CAPS_{870}$C/T* under *MdACS1-1*/2 (**Figure 5.6 C**). There were no significant differences detected between the other allelotypes of $CAPS_{866}$ and $CAPS_{870}$ at a given time point (**Figure 5.6 B**, **Figure 5.7 A-D**). These observations suggested that the direct effect of *MdACS3a* on ethylene production and firmness loss was limited, but its effect on peak ethylene day was clearly detectable through allele *Mdacs3a* (*$CAPS_{870}$T/T*).

The analyses also provided information regarding the effect of *MdACS1* under the same background of $CAPS_{866}$ (**Figure 5.6 B**, **Figure 5.7 A, C**) or $CAPS_{870}$ (**Figure 5.6 C**, **Figure 5.7 B, D**) allelotypes. As expected, allelotype *MdACS1-1/1* had higher ethylene production (**Figure 5.7 A, C**) and more firmness loss (**Figure 5.7 B, D**) than *MdACS1-2/2*, but had similar peak ethylene day as *MdACS1-1/2* (**Figure 5.6 B, C**) except under the *$CAPS_{870}$C/C* background (**Figure 5.6 C**). These results suggested that the effect of *MdACS1* on peak ethylene day was insignificant under the same background of *MdACS3a*, which was in disagreement with the observation that the effect of *MdACS1* on peak ethylene day was significant when the background of *MdACS3a* was not considered (**Figure 5.6 A**).

Since the *MdACS3a* allelotype *CAPS$_{866}$T/T* (*MdACS3a-G289V/G289V*) was present only in four of 97 accessions, the two controlled crosses GMAL4592 and GMAL4593 segregating for *CAPS$_{866}$T/T* under the same background of *MdACS1-1/2* were used for a better analysis. In total, 17 progeny of allelotype *CAPS$_{866}$G/G* (*MdACS3a/MdACS3a*) and another 17 of *CAPS$_{866}$T/T* were similarly evaluated for ethylene production and fruit firmness loss. ANOVA analysis indicated that there were no significant differences between the two allelotypes *CAPS$_{866}$G/G* and *CAPS$_{866}$T/T* in ethylene production and fruit firmness loss, nor in peak ethylene day from d0 through d20 (**Figure S5.2 A-C**), suggesting that no effect of allelotype *CAPS$_{866}$T/T* (*MdACS3a-G289V/G289V*) was detectable in this study.

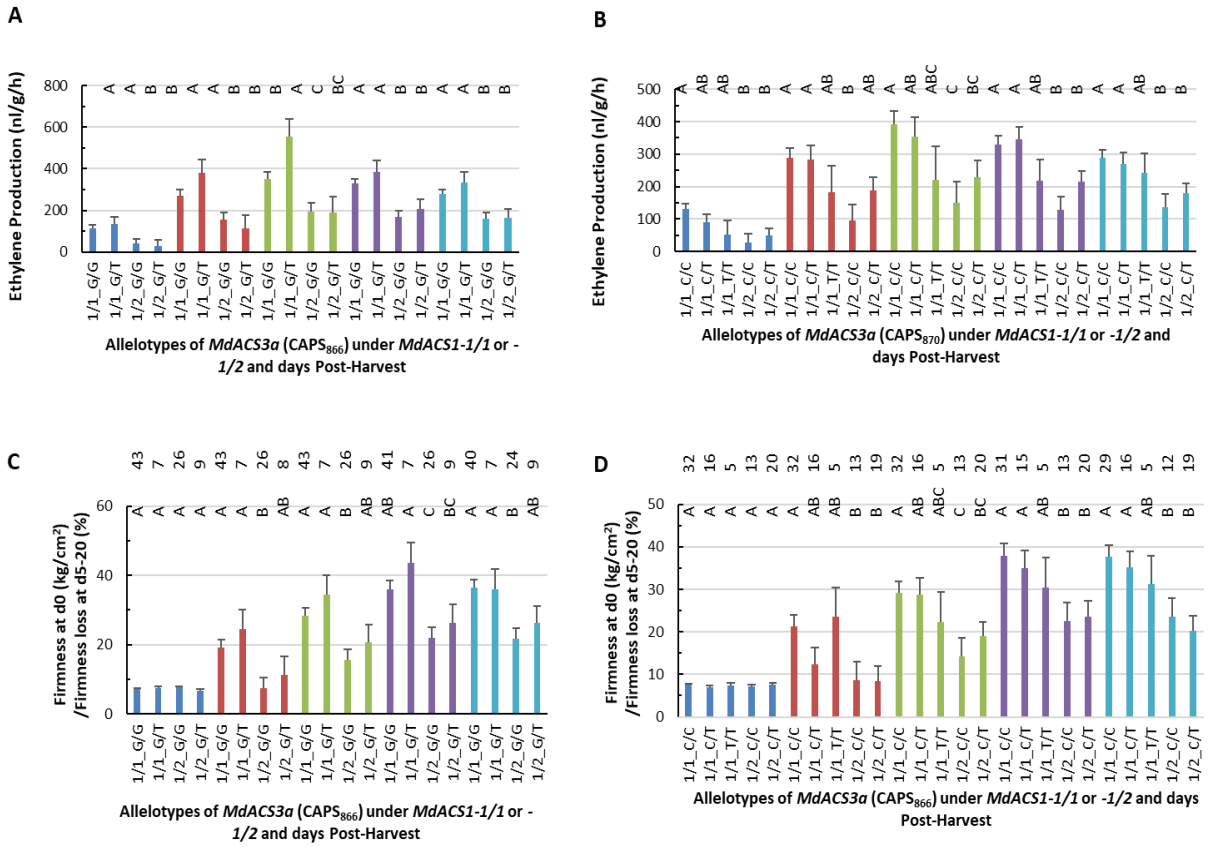


Figure 5.7. Comparison of the means of ethylene production and fruit firmness or firmness loss among the allelotypes of *MdACS3a* as defined by markers CAPS866 (**a, c**) and CAPS870 (**b, d**) under the same background of *MdACS1-1/1* or *MdaCS1-1/2*. The allelotypes, column colors, statistical tests, significance levels and observed numbers are represented similarly as those in Figures 3 and 4.

Allelotyping of *MdACS1* and *MdACS3a* in a large set of *Malus* accessions

Additional 855 *Malus* accessions were surveyed with markers ACS1, $CAPS_{866}$ and $CAPS_{870}$, leading to a total of 952 *Malus* accessions allelotyped (**Figure 5.8**, **Table S5.1**). The data showed that the three allelotypes *MdACS1-1/1*, *MdACS1-1/2* and *MdACS1-2/2* were of 665, 249 and 38 accessions, the allelotypes $CAPS_{866}G/G$, $CAPS_{866}G/T$ and $CAPS_{866}T/T$ were of 770, 173 and 9 accessions, and the allelotypes $CAPS_{870}C/C$, $CAPS_{870}C/T$ and $CAPS_{870}T/T$ were of 346, 400 and 206 accessions, respectively. Estimating the allele frequency in the 952 accessions revealed alleles *MdACS1-1* and *MdACS1-2* of 82.9% and 17.1%, $CAPS_{866}G$ and $CAPS_{866}T$ of 90.0% and 10.0%, and $CAPS_{870}C$ and $CAPS_{870}T$ of 57.4% and 42.6%, respectively (**Figure 5.9 A**).

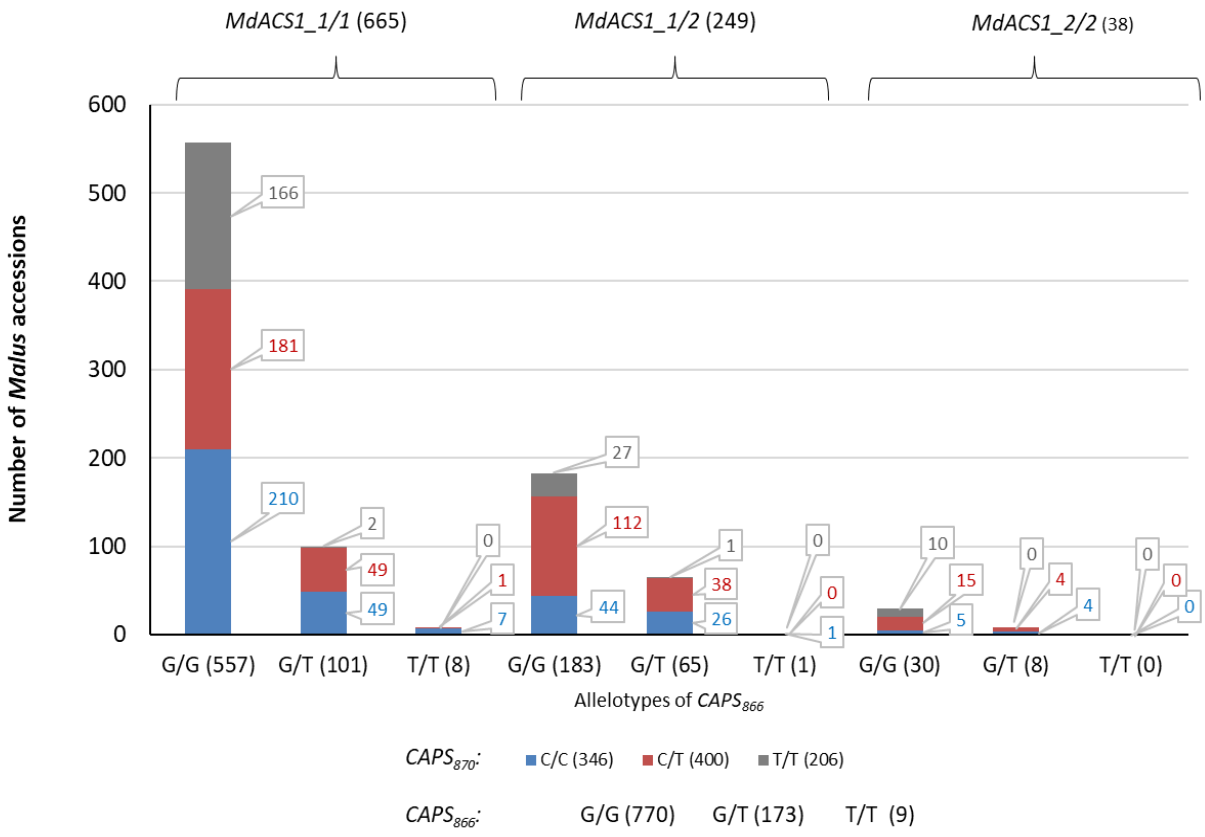


Figure 5.8. Allelotyping of *MdACS1* and *MdACS3a* using markers ACS1, CAPS866 and CAPS870 in 952 *Malus* accessions. The numbers in parentheses stand for the total or subtotal number of *Malus* accessions in an allelotype proximately annotated. The allelotypes are represented similarly as those in Fig.3.

To investigate whether and how human selection might have favored or repressed these alleles, their frequency in the most represented species *M. domestica* (508 accessions), *M. hybrid* (146), and *M. sieversii* (78), which collectively accounted for 76.9% of the 952 accessions (Table S1), were independently estimated (**Figure 5.9 B-D**). In comparison with *M. sieversii*, *M. domestica* and *M. hybrid* showed the largest allele frequency increases for alleles *MdACS1-2* (from 0.6% to 18.8-24.5%) and $CAPS_{870}T$ (from 5.1% to 34.3-48.3%), or decreases for allele *MdACS1-1* (from 99.4% to 81.2-75.5%) and $CAPS_{870}C$ (from 94.9% to 65.7-51.7%), but minimal changes for the $CAPS_{866}G$ (from 86.5% to 86.8-94.2%) and $CAPS_{866}T$ (from 13.5% to 13.2-5.8%) alleles (**Figure 5.9 B-D**). These results suggested that apple breeding practice may

have selected for alleles *MdACS1-2* and *CAPS$_{870}$T* (*Mdacs3a*), against alleles *MdACS1-1* and *CAPS$_{870}$C*, and in neutral for alleles *CAPS$_{866}$G* and *CAPS$_{866}$T* (*MdACS3a-G289V*). Such human selection for alleles *MdACS1-2* and *Mdacs3a* supported their observed significant effect on reduced or delayed ethylene production. Meanwhile, the minimal changes in the frequency of allele *MdACS3a-G289V* reinforced the unfound effect of this allele on ethylene.

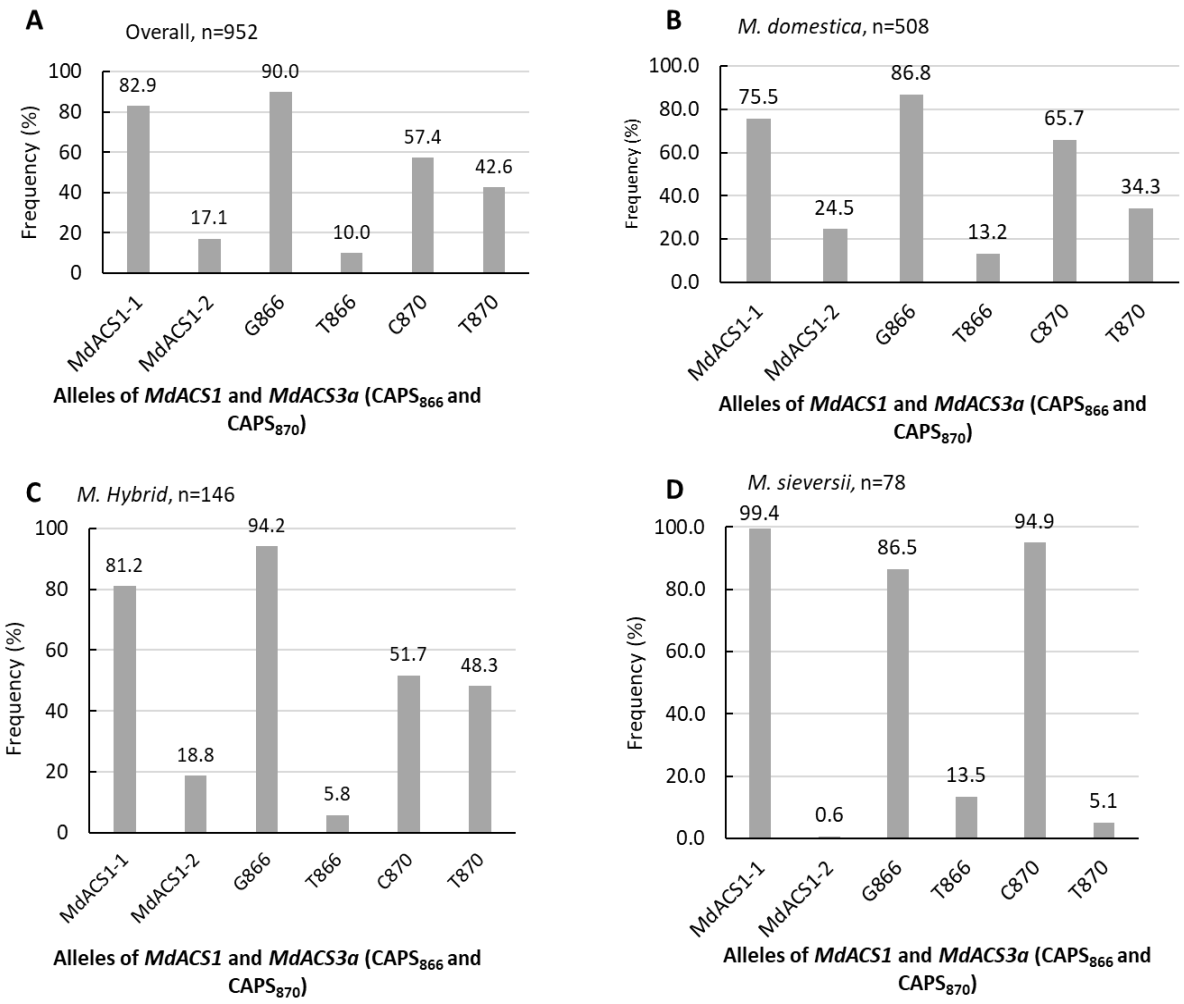


Figure 5.9. Frequency of the *MdACS1* and *MdACS3a* alleles as defined by markers ACS1, CAPS866 and CAPS870 in all the 952 *Malus* accessions (**A**), *M. domestica* (**B**), *M. hybrid* (**C**) and *M. sieversii* (**D**).

Discussion

The effect of *MdACS1* and *MdACS3a* and beneficial alleles

The allelic effect of *MdACS1* on fruit ethylene production and softening was significant and detectable at nearly all time points tested during the 20-d postharvest period in the 97 *Malus* accessions. This was consistent with the critical role of *MdACS1* reported in many other studies (Bulens et al., 2014; Costa et al., 2005; Harada et al., 2000; S. Kondo et al., 2012; Marić &

Lukić, 2014; Hilde Nybom, Ahmadi-Afzadi, Sehic, & Hertog, 2013; H. Nybom et al., 2008; Oraguzie et al., 2004; Sato et al., 2004; Sunako et al., 1999; Zhu & Barritt, 2008). Since the allele frequency of *MdACS1-2* was 24.5% in *M. domestica*, 18.8% in *M. hybrid*, and only 0.6% in *M. sieversii* (**Figure 5.9**), which is the major progenitor species of domestic apples, artificial selection has clearly favored *MdACS1-2* over *MdACS1-1*. In fact, such allele preference of *MdACS1-2* over *MdACS1-1* was even reported within *M. domestica* when the frequencies of the two alleles in apple cultivars were plotted against their time of introduction (H. Nybom et al., 2008). These observations are in accordance with the finding that allele *MdACS1-2* is a beneficial allele associated with low ethylene and slow softening (**Figure 5.5**, **Figure 5.7**).

MdACS3a was regarded a main regulator for ethylene production transition from system 1 to system 2 (Aide Wang et al., 2009). The gene was also similarly shown to be an accelerator (Varanasi, Shin, Mattheis, Rudell, & Zhu, 2011) or an inducer (Busatto et al., 2015) of apple fruit ripening based on its gene expression timing and patterns in apple cultivars of varying ethylene levels and softening rates. In this study, such roles of *MdACS3a* was also detected through examining the allelic effect of *Mdacs3a* (*CAPS$_{870}$T*) on peak ethylene day, which reflects the timing of the climacteric ethylene burst. For example, under the same background of *MdACS1-1*/*1*, allelotype *Mdacs3a*/*Mdacs3a* (*CAPS$_{870}$T/T*) showed a significant delay in peak ethylene day when compared with what was observed for allelotypes *MdACS3a/MdACS3a* (*CAPS$_{870}$G/G*) and *MdACS3a/Mdacs3a* (*CAPS$_{870}$G/T*) (Fig.5c). Moreover, the allele frequency of *Mdacs3a* (*CAPS$_{870}$T*) was 34.3% in *M. domestica* and 48.3% in *M. hybrid*, a dramatic increase from the corresponding frequency of 5.1% in *M. sieversii*, indicating a strong human selection for allele *Mdacs3a*, presumably for the benefit of delayed ethylene production. Taken together, these data support the regulatory role of *MdACS3a* in ethylene production transition in apple

fruit.

However, the allelic effect of *MdACS3a-G289V* on fruit ethylene production, softening and peak ethylene day was shown to be insignificant in the 97 Malus accessions as well as in the 34 progeny from the two controlled crosses segregating for allelotype *MdACS3a-G289V/G289V* (*CAPS$_{866}$T/T*) under the same background of *MdACS1* allelotype. Furthermore, the allele frequency of *MdACS3a-G289V* (*CAPS$_{866}$T*) was 13.5% in *M. sieversii*, 13.2% in *M. domestica* and 5.8% in *M. hybrid*, providing no evidence that *MdACS3a-G289V* (*CAPS$_{866}$T*) has been enriched in response to selection. These results were surprising as *MdACS3a-G289V* was shown to be a functional null allele of *MdACS3a* (Aide Wang et al., 2009). In a previous study, the two null alleles *MdACS3a-G289V* (*CAPS$_{866}$T*) and *Mdacs3a* (*CAPS$_{870}$T*) were concluded to affect the ripening initiation only in late-season apple cultivars, but not in early- or mid-season ones (S. Bai et al., 2012). Such discrepancy in different studies regarding the roles of the two null alleles of *MdACS3a*, particularly *MdACS3a-G289V*, represents a call for more investigations into the role of *MdACS3a-G289V*. Nevertheless, alleles *MdACS1-2* and *Mdacs3a* (*CAPS$_{870}$T*) are clearly demonstrated to be beneficial for breeding apples of low or delayed ethylene profiles in this study, a first effort that simultaneously assessed the roles of *MdACS1* and *MdACS3a* in fruit ethylene production and softening in highly diverse *Malus* materials.

Markers ACS1, CAPS$_{866}$ and CAPS$_{870}$

The assessment of the roles of *MdACS1* and *MdACS3a* in apple fruit ethylene production and softening largely relied on the previously developed marker ACS1 (Harada et al., 2000; Sunako et al., 1999) and the two markers CAPS$_{866}$ and CAPS$_{870}$ developed in this study. Since CAPS$_{866}$ directly detects the mutation SNP G$_{866}$/T$_{866}$, CAPS$_{866}$ is an unequivocal marker for

identifying the functionally null allele *MdACS3a-G289V* (Aide Wang et al., 2009). Marker $CAPS_{870}$ detects SNP C_{870}/T_{870} that does not correspond to a change in the encoding amino acid, i.e. $CAPS_{870}$ detects a silent mutation in *MdACS3a*. Regardless of the nature of SNP C_{870}/T_{870}, T_{870} is a genetic signature for allele *Mdacs3a* as the mutation was identified in 'Fuji', the very source from which the transcriptional null allele *Mdacs3a* was originally defined (Aide Wang et al., 2009). Based on the genomic DNA sequences from 'Fuji', alleles *MdACS3a* (JF833309) and *Mdacs3a* (JF833309) differ by 14 nucleotides, and of these, only four were within the coding sequence (S. Bai et al., 2012). Sequencing of the 92 *Malus* accessions in this study indicated that SNP C_{870}/T_{870} is authentic and varying only between two nucleotides C_{870} and T_{870} (**Figure 5.3**, **Table S5.3**). These data strongly support that $CAPS_{870}$ is a reliable marker for detecting allele *Mdacs3a*. Since both $CAPS_{866}$ and $CAPS_{870}$ detect the characterized SNPs in the coding sequence of *MdACS3a* and can be simply performed by electrophoresis on agarose gels, the two markers are readily applicable for marker assisted selection (MAS) in apple breeding.

Since SNP C_{870}/T_{870} is located only four bases downstream SNP G_{866}/T_{866}, markers $CAPS_{866}$ and $CAPS_{870}$ were once considered to be used as a single marker in this study. But such usage would lead to an ambiguous scenario for allelotype $G_{866}T_{866}/C_{870}T_{870}$ as it could be formed by a combination either between gametes $G_{866}T_{870}$ and $T_{866}C_{870}$ or between gametes $G_{866}C_{870}$ and $T_{866}T_{870}$. To avoid such possible uncertainty, the two markers were used independently.

Previously, an SSR marker targeting at the promoter region of *MdACS3a* was developed and used to allelotype *MdACS3a* in 103 apple varieties (S. Bai et al., 2012). It was shown that three alleles (331bp, 353bp, and 359bp) of the SSR marker corresponded to the wild type allele *MdACS3a* (i.e. *MdACS3a-1* in (S. Bai et al., 2012)), two alleles (333bp and 335bp) to *Mdacs3a* (i.e. *MdACS3a*-2) and one allele (361bp) to *MdACS3a-G289V* (i.e. *MdACS3a-1V*). This makes

the corresponding relationship between the SSR marker alleles and the *MdACS3a* alleles somewhat indirect and inconvenient. Since the size of the SSR marker alleles frequently differ by two base-pairs, an automatic DNA sequencer based detection system is necessary, thereby requiring more sophisticated handling and analysis, compared with the agarose gel based markers $CAPS_{866}$ and $CAPS_{870}$. However, identical allelotypes were observed for all 19 apple cultivars used by co-insistence in both studies (**Table S5.4**), suggesting that the SSR marker and the two CAPS markers are useful for allelotyping of *MdACS3a*. As expected, identical allelotypes for *MdACS1* were also obtained for the 19 common apple cultivars between these two studies (**Table S5.4**).

It should be mentioned that two degenerated CAPS (dCAPS) markers were developed to confirm alleles *Mdacs3a* and *MdACS3a-G289V* in cDNA, but the two dCAPS markers were not used for allelotyping the *MdACS3a* alleles (S. Bai et al., 2012). Therefore, the applicability of the dCAPS markers is unknown in diverse apples.

Utility of the data

Of the 952 *Malus* accessions, 97 were evaluated for their fruit ethylene production and softening at five time points over a 20-d postharvest period (**Table S5.3**). Although most accessions seemed to have predictable ethylene-regulated postharvest behaviors, ‘Virginia Gold’ (PI588778, *M. domestica*) was unusual as it had minimal firmness loss (comparable to ‘Fuji’) during the 20-d storage while producing high levels of ethylene (comparable to ‘Golden Delicious’). This suggested that the slow softening (long shelf life) character of ‘Virginia Gold’ is likely less dependent on ethylene production. More importantly, 'Virginia Gold' has also been shown with an excellent storability (Kamath, Kushad, & Barden, 1992). To understand the lack

of ethylene-related softening in 'Virginia Gold', several preliminary experiments have been initiated by the authors. In melon, it was reported that flesh softening involved both ethylene-dependent and independent components (Pech, Bouzayen, & Latché, 2008). In tomato, the ethylene-independent aspects of fruit ripening was evidenced to be regulated by the FRUITFULL homologs (Bemer et al., 2012). It is possible that investigating fruit softening independent of or less dependent on ethylene production would lead to new knowledge for better understanding the apple fruit ripening process, promising an interesting research area in apple postharvest biology.

In addition, the dataset of allelotypes for genes *MdACS1* and *MdACS3a* generated in the 952 *Malus* accessions would be useful for other future studies involving *MdACS1* and *MdACS3a*, which are the only two apple ACS genes known to be expressed specifically in fruit and associated with apple fruit ethylene production and firmness (Satoru Kondo et al., 2009; Wiersma et al., 2007). The dataset, together with three markers ACS1, $CAPS_{866}$ and $CAPS_{870}$, would be also useful for planning new crosses for developing improved apples with low ethylene and reduced loss of firmness.

Usage of terms allelotype and allelotyping

Term allelotype is defined as "the frequency of alleles in a breeding population." according to *A Dictionary of Genetics* (King, Mulligan, & Stansfield, 2013). In this study, allelotype is referred to the allele composition at a specific gene locus, i.e. *MdACS1* or *MdACS3a*, in individual accessions, highly similar to term 'genotype' for a given DNA marker. Such usage of allelotype represents a drift from or an expansion for the original definition of allelotype defined in the dictionary. But the usage offers convenience for describing allele

composition at a specific gene locus. Indeed, such usage has been adapted already in literature (S. Bai et al., 2012; Sato et al., 2004; Aide Wang et al., 2009).

The definition for term allelotyping in *Encyclopedia of Genetics, Genomics, Proteomics, and Informatics* (Rédei, 2008) reads "Allelotyping is the determination of the spectrum and frequency of allelic variations in a population." The usage of allelotyping in this study is largely covered by the definition, but an extension to include activities for determining allelotype (allele composition at a specific gene locus) is also practiced.

Conclusions

A substantial effort to simultaneously assess the roles of *MdACS1* and *MdACS3a* in fruit ethylene production and softening in diverse *Malus* materials is presented in this study. The most relevant findings include: 1) *MdACS1* had much greater direct influence on fruit ethylene production and softening than *MdACS3a*. 2) Allele *MdACS1-2* was associated with low ethylene and slow softening while *MdACS1-1* with high ethylene and rapid softening. 3) Under the same background of *MdACS1* allelotypes, the transcriptional null allele *Mdacs3a*, rather than the functional null allele *ACS3a-G289V*, significantly delayed the timing to reach climacteric ethylene peak. 4) Alleles *MdACS1-2* and *Mdacs3a* but not *ACS3a-G289V* were highly enriched in *M. domestica* and *M. hybrid* when compared with those in the *M. sieversii*. Overall, this study provides important information as to which alleles of *MdACS1* and *MdACS3a* are beneficial for low and delayed ethylene production and how these beneficial alleles can be selected for apple improvement.

Acknowledgements

This work was financially supported in part by the National Plant Germplasm System (NPGS), Apple Crop Germplasm Committee (CGC), Federal Formula Funds, and College of

Agriculture and Life Science, Cornell University. The final version of this chapter was published in Horticulture Research and can be found at 10.1038/hortres.2016.24. This was co-authored by Yuandi Zhu and Kenong Xu.

REFERENCES

Bai, S., Wang, A., Igarashi, M., Kon, T., Fukasawa-Akada, T., Li, T., Harada, T., & Hatsuyama, Y. (2012). Distribution of MdACS3 null alleles in apple (Malus x domestica Borkh.) and its relevance to the fruit ripening characters. *Breeding Science, 62*(1), 46-52.

Bai, Y., Dougherty, L., Li, M., Fazio, G., Cheng, L., & Xu, K. (2012). A natural mutation-led truncation in one of the two aluminum-activated malate transporter-like genes at the Ma locus is associated with low fruit acidity in apple. *Mol Genet Genomics, 287*(8), 663-678.

Barry, C. S., & Giovannoni, J. J. (2007). Ethylene and fruit ripening. *Journal of Plant Growth Regulation, 26*(2), 143-159.

Bemer, M., Karlova, R., Ballester, A. R., Tikunov, Y. M., Bovy, A. G., Wolters-Arts, M., Rossetto, P. d. B., Angenent, G. C., & de Maagd, R. A. (2012). The Tomato FRUITFULL Homologs TDR4/FUL1 and MBP7/FUL2 Regulate Ethylene-Independent Aspects of Fruit Ripening. *The Plant Cell, 24*(11), 4437-4451.

Blanpied, G. D., & Silsby, K. J. (1992). Predicting harvest date windows for apples. In *Information Bulletin 221*: Cornell Cooperative Extension, Cornell University, Ithaca, NY, USA.

Boualem, A., Fergany, M., Fernandez, R., Troadec, C., Martin, A., Morin, H., Sari, M. A., Collin, F., Flowers, J. M., Pitrat, M., Purugganan, M. D., Dogimont, C., & Bendahmane, A. (2008). A conserved mutation in an ethylene biosynthesis enzyme leads to andromonoecy in melons. *Science, 321*(5890), 836-838.

Bulens, I., Van de Poel, B., Hertog, M. L. A. T. M., Cristescu, S. M., Harren, F. J. M., De Proft, M. P., Geeraerd, A. H., & Nicolai, B. M. (2014). Dynamic changes of the ethylene biosynthesis in 'Jonagold' apple. *Physiologia Plantarum, 150*(2), 161-173.

Burmeister, D. M., & Dilley, D. R. (1995). A 'scald-like' controlled atmosphere storage disorder of Empire apples -- a chilling injury induced by CO2. *Postharvest Biology and Technology, 6*(1-2), 1-7.

Busatto, N., Farneti, B., Tadiello, A., Velasco, R., Costa, G., & Costa, F. (2015). Candidate gene expression profiling reveals a time specific activation among different harvesting dates in 'Golden Delicious' and 'Fuji'apple cultivars. *Euphytica*, 1-13 (First online: 15 December 2015).

Cara, B., & Giovannoni, J. J. (2008). Molecular biology of ethylene during tomato fruit development and maturation. *Plant Science, 175*(1-2), 106-113.

Costa, F., Stella, S., Van de Weg, W. E., Guerra, W., Cecchinel, M., Dalla Via, J., Koller, B., & Sansavini, S. (2005). Role of the genes Md-ACO1 and Md-ACS1 in ethylene production and shelf life of apple (Malus domestica Borkh). *Euphytica, 141*(1-2), 181-190.

Fawbush, F., Nock, J. F., & Watkins, C. B. (2008). External carbon dioxide injury and 1-methylcyclopropene (1-MCP) in the 'Empire' apple. *Postharvest Biology and Technology, 48*(1), 92-98.

Forsline, P. L., Aldwinckle, H. S., Dickson, E. E., Luby, J. J., & Hokanson, S. C. (2003). Collection, Maintenance, Characterization and Utilization of Wild Apples of Central Asia. *Horticultural Reviews, 29* 1-61.

Harada, T., Sunako, T., Wakasa, Y., Soejima, J., Satoh, T., & Niizeki, M. (2000). An allele of the 1-aminocyclopropane-1-carboxylate synthase gene (Md-ACS1) accounts for the low level of ethylene production in climacteric fruits of some apple cultivars. *Theoretical and Applied Genetics, 101*(5-6), 742-746.

Kamath, O. C., Kushad, M. M., & Barden, J. A. (1992). Postharvest Quality of 'Virginia Gold' Apple Fruit. . *Fruit Var J.*

, *46*, 87-92.

King, R. C., Mulligan, P., & Stansfield, W. (2013). *A dictionary of genetics (8th Edition)*: Oxford University Press.

Kondo, S., Meemak, S., Ban, Y., Moriguchi, T., & Harada, T. (2009). Effects of auxin and jasmonates on 1-aminocyclopropane-1-carboxylate (ACC) synthase and ACC oxidase gene expression during ripening of apple fruit. *Postharvest Biology and Technology, 51*(2), 281-284.

Kondo, S., Tomiyama, H., Kittikorn, M., Okawa, K., Ohara, H., Yokoyama, M., Ifuku, O., Saito, T., Ban, Y., Tatsuki, M., Moriguchi, T., Murata, A., & Watanabe, N. (2012). Ethylene production and 1-aminocyclopropane-1-carboxylate (ACC) synthase and ACC oxidase gene expression in apple fruit are affected by 9,10-ketol-octadecadienoic acid (KODA). *Postharvest Biology and Technology, 72*, 20-26.

Li, T., Tan, D., Yang, X., & Wang, A. (2013). Exploring the apple genome reveals six ACC synthase genes expressed during fruit ripening. Scientia Horticulturae, 157, 119-123. doi:https://doi.org/10.1016/j.scienta.2013.04.016

Marić, S., & Lukić, M. (2014). Allelic polymorphism and inheritance of MdACS1 and MdACO1 genes in apple (Malus × domestica Borkh.). *Plant Breeding, 133*(1), 108-114.

Neff, M. M., Turk, E., & Kalishman, M. (2002). Web-based primer design for single nucleotide polymorphism analysis. *Trends in Genetics, 18*(12), 613-615.

Nybom, H., Ahmadi-Afzadi, M., Sehic, J., & Hertog, M. (2013). DNA marker-assisted evaluation of fruit firmness at harvest and post-harvest fruit softening in a diverse apple germplasm. *Tree Genetics & Genomes, 9*(1), 279-290.

Nybom, H., Sehic, J., & Garkava-Gustavsson, L. (2008). Modern apple breeding is associated with a significant change in the allelic ratio of the ethylene production gene Md-ACS1. *Journal of Horticultural Science & Biotechnology, 83*(5), 673-677.

Oraguzie, N. C., Iwanami, H., Soejima, J., Harada, T., & Hall, A. (2004). Inheritance of the Md-ACS1 gene and its relationship to fruit softening in apple (Malus x domestica Borkh.). *Theoretical and Applied Genetics, 108*(8), 1526-1533.

Pech, J. C., Bouzayen, M., & Latché, A. (2008). Climacteric fruit ripening: Ethylene-dependent and independent regulation of ripening pathways in melon fruit. *Plant Science, 175*(1–2), 114-120.

Rédei, G. P. (2008). *Encyclopedia of genetics, genomics, proteomics, and informatics (3rd Edition)* (3rd ed.): Springer Science & Business Media.

Rosenfield, C.-L., Kiss, E., & Hrazdina, G. (1996). MdACS-2 (Accession No. U73815) and MdACS-3 (Accession No. U73816): Two New l-Aminocyclopropane-1-Carboxylate Synthases in Ripening Apple Fruit. *Plant Physiol. 112 (4), 1735-1736 (1996), 112*(4), 1735.

Sato, T., Kudo, T., Akada, T., Wakasa, Y., Niizeki, M., & Harada, T. (2004). Allelotype of a ripening-specific 1-aminocyclopropane-1-carboxylate synthase gene defines the rate of fruit drop in apple. *Journal of the American Society for Horticultural Science, 129*(1), 32-36.

Singh, V., Weksler, A., & Friedman, H. (2017). Different Preclimacteric Events in Apple Cultivars with Modified Ripening Physiology. Frontiers in Plant Science, 8, 1502-1502. doi:10.3389/fpls.2017.01502

Sunako, R., Sakuraba, W., Senda, M., Akada, S., Ishikawa, R., Niizeki, M., & Harada, T. (1999). An allele of the ripening-specific 1-aminocyclopropane-1-carboxylic acid synthase gene (ACS1) in apple fruit with a long storage life. *Plant Physiology, 119*(4), 1297-1303.

Varanasi, V., Shin, S., Mattheis, J., Rudell, D., & Zhu, Y. (2011). Expression profiles of the MdACS3 gene suggest a function as an accelerator of apple (Malus × domestica) fruit ripening. *Postharvest Biology and Technology, 62*(2), 141-148.

Wang, A., & Xu, K. (2012). Characterization of Two Orthologs of REVERSION-TO-ETHYLENE SENSITIVITY1 in Apple. *Journal of Molecular Biology Research, 2*(1), 24-41.

Wang, A., Yamakake, J., Kudo, H., Wakasa, Y., Hatsuyama, Y., Igarashi, M., Kasai, A., Li, T., & Harada, T. (2009). Null mutation of the MdACS3 gene, coding for a ripening-specific 1-aminocyclopropane-1-carboxylate synthase, leads to long shelf life in apple fruit. *Plant physiology, 151*(1), 391-399.

Watkins, C. B., Erkan, M., Nock, J. F., Iungerman, K. A., Beaudry, R. M., & Moran, R. E. (2005). Harvest date effects on maturity, quality, and storage disorders of 'Honeycrisp' apples. *Hortscience, 40*(1), 164-169.

Watkins, C. B., Silsby, K. J., & Goffinet, M. C. (1997). Controlled Atmosphere and Antioxidant Effects on External CO2 Injury of `Empire' Apples. *Hortscience, 32*(7), 1242-1246.

Wiersma, P. A., Zhang, H., Lu, C., Quail, A., & Toivonen, P. M. A. (2007). Survey of the expression of genes for ethylene synthesis and perception during maturation and ripening

of 'Sunrise' and 'Golden Delicious' apple fruit. *Postharvest Biology and Technology*, *44*(3), 204-211.

Yang, S. F., & Hoffman, N. E. (1984). Ethylene Biosynthesis and Its Regulation in Higher-Plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, *35*, 155-189.

Zhu, Y., & Barritt, B. (2008). Md-ACS1 and Md-ACO1 genotyping of apple (Malus x domestica Borkh.) breeding parents and suitability for marker-assisted selection. *Tree Genetics & Genomes*, *4*(3), 555-562.

Supplementary Figures

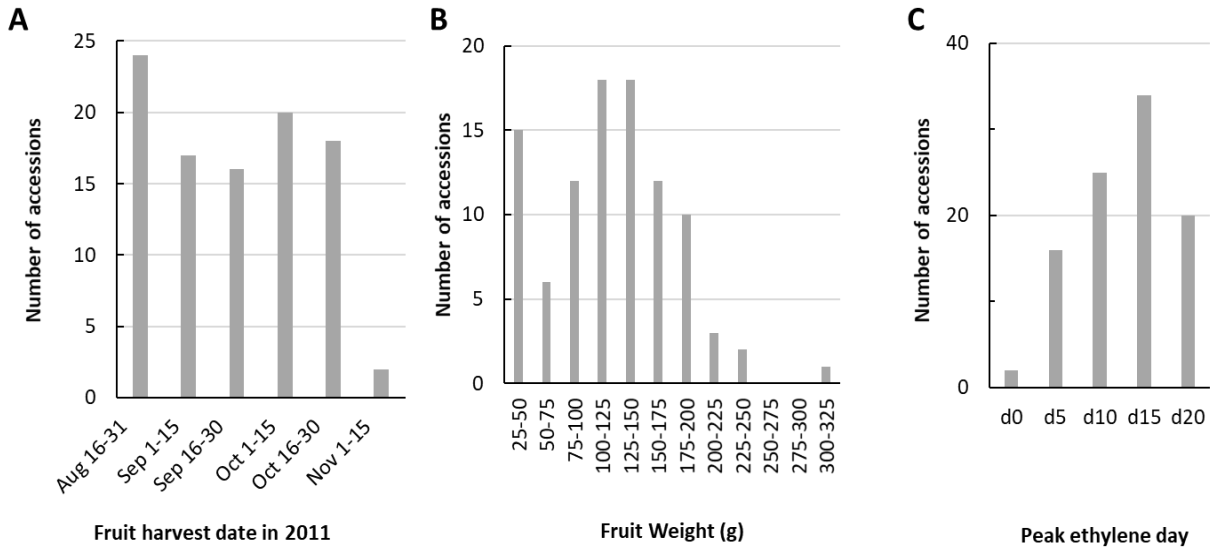


Figure S5.1. Distribution of fruit maturity/harvest date (**A**), fruit weight (**B**), and peak ethylene day (**C**).

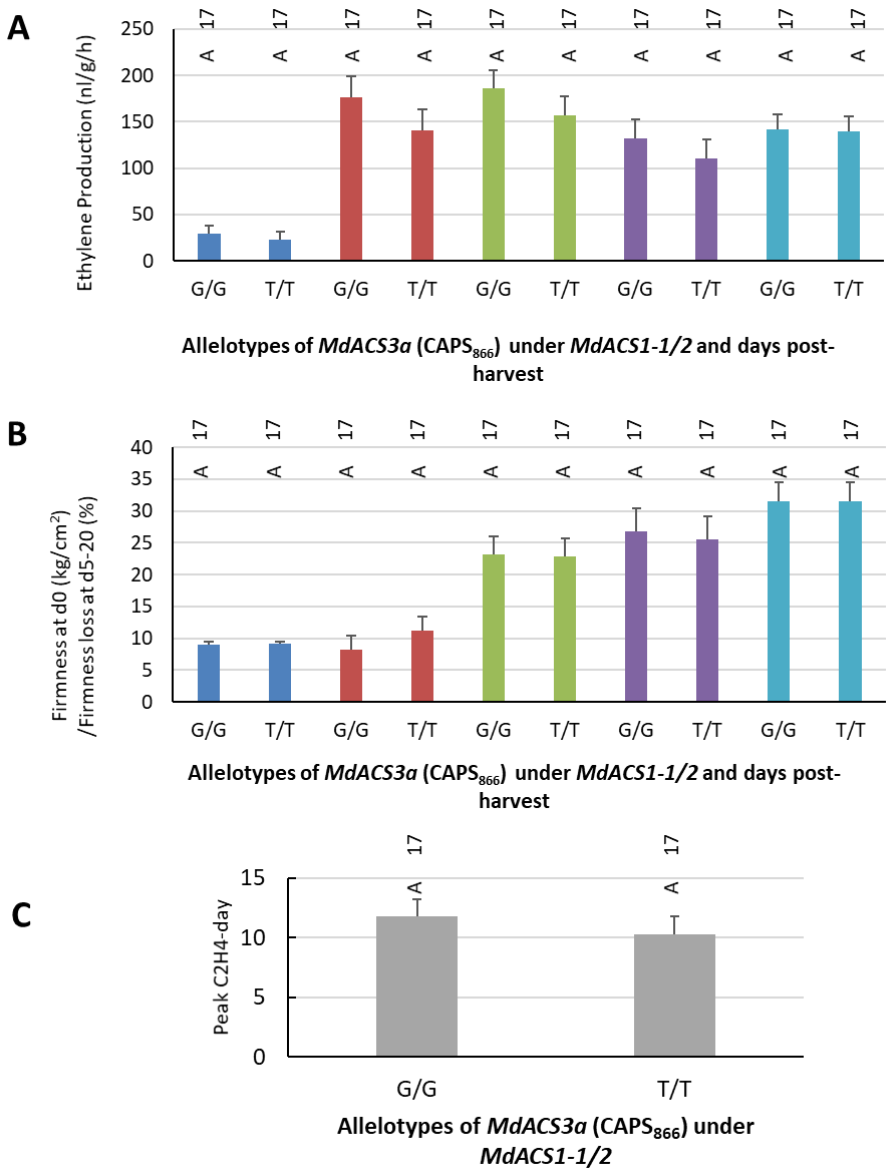


Figure S5.2. Comparison of the means of ethylene production (**A**), fruit firmness (**B**) and peak ethylene day (**C**) among allelotypes of *MdACS3a* as defined by marker CAPS866 under the same background of *MdACS1-1/2* in the 34 progeny of crosses GMAL4592 and GMAL4593. Allelotypes *MdACS3a*/*MdACS3a* (G_{866}/G_{866}) and *MdACS3a-G289V/G289V* (T_{866}/T_{866}) are noted with 'G/G' and 'T/T', respectively. Colors of column in blue, orange, green, purple and turquoise represent d0, d5, d10, d15 and d20, respectively. Significance levels are indicated with letters, where different letters indicate $p<0.05$. The numbers of accessions observed (n) for each allelotype are presented accordingly. Error bars indicate standard errors.

Table S5.1. List of 952 *Malu*s accessions allelotyped with markers ACS1 and $CAPS_{866}$ and $CAPS_{870}$.

Table is prodived as a supplementary spreadsheet.

Table S5.2. Allele specific primers for genes *MdACS1* and *MdACS3a*

Gene/Primer Name	Primer Sequence (5’ to 3’)
MdACS1/ACS1-5F	AGAGAGATGCCATTTTTGTTCGTAC
MdACS1/ACS1-5R	CCTACAAACTTGCGTGGGGATTATAAGTGT
MdACS3a/ACS3a-289F	CTTCCAGATTACTCCTCAAGCTTTA
*MdACS3b**	GTTCCAGATTACTCCAGAAGCGTTA
*MdACS3c**	TTTCCAAATCACTCCTCAAGCTTTG
MdACS3a/ACS3a-289R	AGTCTCTTTCTATTTGTCTTTATGTAGTTTC
*MdACS3b**	AGTCTCTCTCTCTGTTTGTATTTATGTAATTTT
*MdACS3c**	AGTCTCTCTCTATTTGTCTTTATGTAATTTT

*The DNA bases of *MdACS3b* and *MdACS3c* highlighted in red show how they are discriminated from *MdACS3a* in the region covered by primers ACS3a-289F/R.

Table S5.3. Evaluation of fruit ethylene production and firmness in a subset of 97 *Malus* accessions

Accession #	Name	Species	Starch Index at Harvest	Fruit Weight-Mean (g)	Harvest date	MdACS1 allelotype	MdACS3a-CAPS$_{8866}$	MdACS3a-CAPS$_{8700}$	C2H4-d0 (nl/g/h)	C2H4-d5 (nl/g/h)	C2H4-d10 (nl/g/h)	C2H4-d15 (nl/g/h)	C2H4-d20 (nl/g/h)	Peak C2H4-day	Firmness d0 (kg)	Firmness lost_d5 (%)	Firmness lost_d10 (%)	Firmness lost_d15 (%)	Firmness lost_d20 (%)
PI123989	Emilia	domestica	7.3	110.1	10/25/2011	1/1	G/G	C/C	101.9	294.9	405.3	362.3	358.4	10.0	6.2	5.77	15.12	23.09	30.24
PI246464	James Grieve (Red Rosamund strain)	domestica	4.0	140.7	8/19/2011	1/1	G/G	T/T	21.1	269.2	358.5	219.0	254.4	10.0	7.4	42.27	44.50	60.31	47.34
PI264693	Sumatovka	domestica	4.3	94.4	10/19/2012	1/1	T/T	C/C	1.2	8.6	40.6	17.1	84.2	20.0	9.3	0.7	2.6	17.4	24.3
PI280400	Anna	domestica	5.8	122.0	9/8/2011	½	G/G	C/T	234.6	223.1	206.8	149.5	168.6	1.0	3.9	8.93	21.60	24.97	28.33
PI323617		pumila	7.8	33.7	8/19/2011	1/1	G/G	C/C	193.2	318.8	382.9	549.9	793.0	20.0	5.5	66.41	47.35	46.72	39.36
PI392303	Gala	domestica	4.4	128.1	9/13/2011	2/2	G/T	C/C	22.1	45.2	161.2	356.6	192.6	15.0	9.7	12.78	28.97	32.22	40.82
PI483257	Reinette Simirenko	domestica	5.9	169.0	10/28/2011	½	G/T	C/T	2.6	64.6	72.7	81.9	93.2	20.0	7.1	-1.12	18.42	27.03	14.71
PI588747	Florina	domestica	7.0	162.9	10/10/2011	½	G/G	C/C	4.3	91.7	255.9	173.2	216.4	10.0	7.9	5.66	9.98	23.91	24.60
PI588772	Monroe	domestica	6.5	148.0	9/26/2011	2/2	G/G	C/T	2.2	10.2	57.3	76.0	57.7	15.0	7.9	-2.96	5.03	10.82	21.02
PI588778	Virginia gold	domestica	6.0	146.9	10/25/2011	1/1	G/G	C/T	14.7	188.8	322.6		205.5	10.0	6.5	0.54	0.38		-2.76
PI588785	Esopus Spitzenburg	domestica	7.0	89.1	10/18/2011	½	G/G	C/T	105.2	248.3	284.2	218.0	263.0	10.0	8.9	-1.07	4.61	17.76	19.28
PI588798	Rambo-Red Summer	hybrid	4.3	239.7	9/17/2011	1/1	G/T	C/T	45.4	484.5	678.9	454.1	406.2	10.0	5.7	30.04	38.86	36.94	42.88
PI588835	Burgundy	domestica	6.0	161.4	8/31/2011	½	G/G	C/C	66.0	115.2	124.4	133.1	68.7	15.0	7.0	2.42	17.09	22.66	34.47

Table S5.3 continued

Accession #	Name	Species	Starch Index at Harvest	Fruit Weight-Mean (g)	Harvest date	MdACS1 allelotype	MdACS3a-CAPS$_{866}$	MdACS3a-CAPS$_{870}$	C2H4-d0 (nl/g/h)	C2H4-d5 (nl/g/h)	C2H4-d10 (nl/g/h)	C2H4-d15 (nl/g/h)	C2H4-d20 (nl/g/h)	Peak C2H4-day	Firmness d0 (kg)	Firmness lost_d5 (%)	Firmness lost_d10 (%)	Firmness lost_d15 (%)	Firmness lost_d20 (%)
PI588837	Gravenstein Washington Red	domestica	5.8	165.1	8/31/2011	1/1	G/G	C/T	40.9	209.4	358.2	299.3	156.2	10.0	5.9	26.99	34.55	46.10	41.60
PI588838	Nova Easygro	domestica	4.4	135.5	9/26/2011	1/1	G/G	T/T	47.1	56.7	62.3	77.5	66.6	15.0	7.8	1.99	9.06	6.23	12.91
PI588841	Idared	domestica	4.6	169.3	10/13/2011	1/2	G/G	C/C	11.1	6.6	34.6	60.6	50.7	15.0	6.6	-2.19	11.49	28.04	31.29
PI588842	Empire	domestica	6.0	103.9	10/10/2011	1/2	G/G	C/C	1.0	2.2	146.5	167.7	175.1	20.0	7.8	10.75	1.79	25.85	36.47
PI588844	Fuji Red Sport Type 2	domestica	7.0	185.5	10/28/2011	2/2	G/G	C/T	0.7	4.7	21.0	49.9	49.7	15.0	7.1	0.58	0.51	1.72	1.47
PI588850	Rome Beauty Law	hybrid	5.2	197.5	10/25/2011	1/2	G/G	T/T	4.7	85.6	136.0	131.9	146.7	20.0	8.2	10.89	15.63	21.53	35.40
PI588872	Northern Spy	domestica		151.9	10/18/2011	1/1	G/G	C/C	280.4	447.6	447.6			5.0	5.4	-10.24	1.00		
PI588880	Granny Smith	domestica	5.6	152.1	11/8/2011	1/2	G/G	C/C	0.8	16.1	90.6	20.0		10.0	7.7	-5.67	-12.98	-8.57	
PI588883	Demir	hybrid	5.2	82.0	10/25/2011	1/1	G/G	C/C	2.1	3.2	66.9	258.0	231.7	15.0	9.6	3.91	-1.67	-1.95	0.97
PI588943	Liberty	domestica	5.0	146.1	10/7/2011	1/1	G/G	C/T	74.9	368.9	509.9	386.7	291.7	10.0	7.6	-1.97	26.18	26.78	25.53

Table S5.4. Comparison of the *MdACS1* and *MdACS3a* allelotypes in *Malus* accessions used in both Bai et al. (Bai et al., 2012) and this study.

Malus accession #	Name	SPECIES	MdACS1 allelotype	MdACS3a-CAPS866	MdACS3a-CAPS870	MdACS1*	MdACS3a-SSR*
PI199525	Amanishiki	domestica	1/2	G/G	C/T	1/2	333/331
PI255899	Akane	domestica	2/2	G/G	C/T	2/2	333/359
PI392303	Gala	domestica	2/2	G/T	C/C	2/2	361/331
PI483255	Priam	domestica	1/2	G/T	C/T	1/2	333/361
PI588785	Esopus Spitzenburg	domestica	1/2	G/G	C/T	1/2	333/331
PI588817	McIntosh	domestica	1/1	G/G	C/T	1/1	335/331
PI588819	Vista Bella	domestica	1/1	G/G	C/T	1/1	333/331
PI588841	Idared	domestica	1/2	G/G	C/C	1/2	353/353
PI588844	Fuji	domestica	2/2	G/G	C/T	2/2	333/331
PI588850	Rome Beauty Law	hybrid	1/2	G/G	T/T	1/2	333/333
PI588853	Cox's Orange Pippin	domestica	1/2	G/G	C/T	1/2	333/331
PI588863	Jerseymac	domestica	1/1	G/G	C/T	1/1	333/331
PI588880	Granny Smith	domestica	1/2	G/G	C/C	1/2	331/331
PI588942	Julyred	domestica	1/1	G/G	C/T	1/1	333/359
PI589067	Redgold	domestica	1/2	G/T	C/C	1/2	361/331
PI589181	Prima	domestica	1/2	G/T	C/T	1/2	333/361
PI589841	Delicious	domestica	1/2	G/G	C/C	1/2	331/331
PI590184	Golden Delicious	domestica	1/2	G/T	C/T	1/2	333/361
PI590185	Jonathan	domestica	1/2	G/G	C/T	1/2	333/353

* From Bai et al. (2012a)

Bai, S., Wang, A., Igarashi, M., Kon, T., Fukasawa-Akada, T., Li, T., Harada, T., & Hatsuyama, Y. (2012). Distribution of MdACS3 null alleles in apple (Malus x domestica Borkh.) and its relevance to the fruit ripening characters. *Breeding Science, 62*(1), 46-52.

CHAPTER 6

Conclusion

Apples are an important fruit crop worldwide. The introduction of new apple cultivars with superior taste, texture and overall quality excites the market. Breeders strive to create the next 'big' apple with grower friendly traits and consumers expected fruit quality. The identification of genes responsible for key traits are critical for informed breeding of these new superior varieties. The research work presented in chapters 2-5 included the initial genetic mapping of the *W (*weeping), identification of *MdLazy1*, mapping of *c2* and *c3* as genetic repressors of columnar growth habit and assessment of *MdACS1* and MdACS3a alleles on fruit ethylene production and fruit softening. These efforts contribute to the overall goal of creating new cultivars by identifying genes responsible for key fruit and tree architecture traits, increasing our understanding of the traits and developing genetic markers breeders can utilize for those traits.

The weeping growth habit is a desired ornamental tree form. The defining downward growing branches are easily manipulated making them attractive for different training system. In chapter 2 *W,* the major locus linked to weeping was identified on chromosome 13 along with three minor loci (*W2, W3* and *W4*) located on chromosomes 10, 16 and 5, respectively. Further investigation of the *W* locus led to the identification of *MdLazy1* as a strong candidate gene for *W.* Characterization of *MdLazy1* revealed two alleles. The *MdLazy1W* allele contains a non-synonymous mutation of L195P that is tightly linked to the weeping phenotype and may be causal. Transgenic trees overexpressing *MdLazy1W* had a weeping like appearance that was similar to transgenic trees with reduced expression of *MdLazy1S.* In the future, gene editing targeting *MdLazy1* in existing varieties could create new tree architectures for growers, while

leaving the fruit unaltered.

The columnar growth habit, characterized by compact growth, short internode length and reduced branching, was first identified as a somatic mutation of ‘McIntosh’ and called ‘Wijcik McIntosh’. A retroposon insertion on chromosome 10 and upregulated expression of a nearby gene called *Co* (columnar) in ‘Wijcik’ but not in ‘McIntosh was reported as the causal factors for the phenotype. In chapter 4, we identified two genetic loci *c2* and *c3* on chromosomes 10 and 9, respectively, which can repress the columnar growth habit in trees containing the retroposon insertion by reducing *Co* expression. Future work is need to identify genes under the *c2* and *c3* loci that are responsible for repressing the columnar phenotype.

Peak harvest season for apples is from August through October in the U.S, yet apples are available year round. Post-harvest storage is critical for year round availability, however some apple varieties store better than others. Poor storage varieties soften or develop physiological disorders over time in storage, limiting their shelf life to harvest season only. Good storage apples maintain their firmness, texture and quality ensuring consumers get a consistent product year round.*MdACS1* and *MdACS3a* are important genes for ethylene production and fruit ripening. In Chapter 5 we developed two new CAPs markers to distinguish two null alleles of *MdACS3a.* We then allelotyped 97 diverse cultivars and 34 progeny from a cross with *MdACS1* and *MdACS3a* markers and evaluated their ethylene production and softening for a 20 day post-harvest period. We determined that the effects of *MdACS1* are more important than those of *MdACS3a* for fruit ethylene production and fruit softening. The best allele combinations for reduced ethylene production, limited softening and therefore better storage were *MdACS1-2* and *Mdacs3a.* This information is expected to helpful for developing new and improved varieties.

The genetic mechanisms underlying the traits investigated in Chapters 2-5 are fascinating

and complex. This work explored weeping and columnar growth habits and apple fruit ethylene production and softening for post-harvest storage. These studies contribute to the overall understanding of tree architecture and fruit quality in apple while posing new questions to be answered.