Identification of Unknown Apple (*Malus* × *domestica*) Varieties in the Duluth Region Using DNA Simple Sequence Repeats

Marshall J Wedger
Swenson College of Science and Engineering
University of Minnesota Duluth
Wedge015@d.umn.edu

Introduction:

There are hundreds of varieties of domesticated apple (*Malus* × *domestica*) grown around the world in temperate regions of the northern and southern hemispheres, from southern United States up through Canada, and from South Africa down into New Zealand. The domesticated apple as we know it is believed to be most closely related to the wild species *M. sieversii* in central Asia (Cornille et al. 2012). The same study reported that the apple spread through North Africa and Europe after its initial domestication, acquiring some genomic contributions from other closely related species, including *M. baccata, M. orientalis,* and *M. sylvetris*. Since the domesticated apple is a long-lived perennial, and therefore has gone through fewer generations than those seen in annuals like wheat and maize, it shows less dramatic changes under domestication (remains highly diverse); however it still has evolved larger and sweeter fruits than its wild counterparts.

Here in Minnesota, and more specifically Duluth, apples are a common part of the landscape; many abandoned trees grow in backyards and public spaces. However, most of the information regarding the origin of these trees is absent, and the specific variety of most trees is unknown. It is commonly speculated that many of the local trees are derived from the UMD trial orchard (active between the early 1910s and the late 1960s), and that a few varieties dominate because of their fitness in our severe climate (C. Hale, personal communication). The UMD Sustainable Agriculture Project (SAP) is currently working to restore the newly rediscovered UMD trial orchard, and the Gross lab (UMD Biology Department) is cooperating in a new project to genotypically identify unknown apple trees in the region. Apples are mostly propagated through grafting, or cloning, which results in genetically identical trees. Thus, this mode of reproduction allows unknown trees to be genetically matched to known, named samples in either local or national collections. The genetic composition of cloned apple trees in the Duluth region has not been tested, despite a strong community interest in tree identification and common speculation about the origin of these trees. I will be testing the null hypothesis of equal representation of each known variety of apple in the Duluth region.

Genotyping unknown apple trees using DNA Simple Sequence Repeats (SSRs) is a way to easily and economically genotype fairly large populations. Using nine previously published SSRs (Gross et al. 2012), and a 'probability of identity' (PI) statistical test as described by Peakall et al. (2006), one can determine with near certainty the correct classification, or "DNA fingerprint" of each individual. The PI test is necessary to ensure that the SSRs are sufficiently variable to reliably separate samples given the size of the dataset, allowing us to essentially eliminate the possibility that two genotypes are identical due to chance alone. A sufficiently low PI value, ensures that apple genotypes are identical because of the clonal mode of propagation used in cultivated settings. Previous studies (Hokanson et al., 1998, Gross et al. 2012) have indicated that the SSRs to be used in this project are sufficiently variable to ensure a low enough PI value; for example, Hokanson et al. (1998) calculated that in their sample set, the probability of two unrelated plants having the same genotypes at all SSRs by chance is less than 1 in 1 billion. Identification of Duluth area trees was facilitated by cross-referencing the local genotypes with over 1,000 named, genotyped cultivars from the USDA National Plant Germplasm System (USDA-NPGS) apple collection.

Materials and Methods:

Sample Collection and DNA Preparation

Malus samples for this study were collected from multiple locations. Thirty-four samples were collected from the UMD trial orchard, 37 were collected from community orchards, and 90 were collected from community members via farmer's markets. Twenty-one standards were also ordered through the USDA-NPGS orchard. A total of 184 samples were collected. After collection, DNA was extracted using the Fermentas GeneJET™ Plant Genomic DNA Purification Mini Kit, following protocol given in the manual. DNA was diluted to 2ng/µL for downstream applications.

Genetic Markers

Nine previously published (Gross et al. 2012) primers were used (GD12, GD15, GD96, GD142, GD147, GD162, CH01h01, CH01f02, and CH02d08). These primers were selected for their separation of allele sizes, their high levels of variation, and overlap with previous studies using SSRs to genotype apple trees in other countries. A three primer system (Schuelke 2000) was used in order to fluorescently label polymerase chain reaction (PCR) products. Standard reverse primers were used, along with a special forward primer with generic tail that attached to a fluorescent, generic primer (FAM, NED, PET). This was done so the products could be visualized on the ABI 3730 capillary sequencer. The PCR program consisted of a touchdown protocol. An initial denaturing cycle of 2 min at 95 ^{o}C was followed by 10 touchdown cycles (the

program starts 10 ^{o}C above the appropriate annealing temperature, and the temperature drops 1 ^{o}C each cycle) of 30 s at 94 ^{o}C, 30 s at the annealing temperature (Table 1.), and 1 min. at 72 ^{o}C. These 10 cycles were followed by 29 cycles of 30 s at 94 ^{o}C, 30 s at the appropriate annealing temperature, and 1 min. at 72 ^{o}C, with a final elongation period at 72 ^{o}C for 10 min. After PCR amplification the samples were multiplexed as shown in Table 1 and sent to the University of Minnesota-Twin Cities Genomics Center (UMGC) to be visualized on the ABI 3730 capillary sequencer.

Data Analysis

Results returned from the UMGC were scored by hand using the GeneMarker V2.4.0 software. After allele scoring, samples of known allele size were used to standardize allele sizes to USDA-NPGS values. Only samples with zero or one markers showing polyploid status were used in downstream applications (N=177). With these standardized values and 1910 values from previous work (Gross 2012), a Probability of Identity test was performed using GenAlEx 6.501 as described in Peakall et al. (2006). The data collected in this experiment was further cut to only those samples that had missing data at one or zero loci (missing from PCR testing or polyploidy status), or were not previously known (standards). The data collected in this experiment (N'=140) was then compared to the USDA-NPGS collection to look for matching genotypes. To achieve this GENODIVE V2.0b23 (Meirmans and Van Tienderen 2004) software was used at a threshold value T=2, which allows one scoring error across all nine loci. Allele frequencies and genetic diversity statistics were also calculated with GENODIVE.

Results:

Data Treatment

A total of 184 samples were genotyped, and after analysis with GeneMarker only samples with data missing for a maximum of one loci were retained, reducing the sample size to 140. Using the genetic data from these 140 samples and another 1,910 samples from previous research (Gross et al. 2012) GenAlEx returned a PI value of $5.2X10^{-14}$, or about 1 in 19 trillion possibility that any matching genotype is due to chance alone, for the standard probability of identity. For the more conservative PIsibs, GenAlEx returned a value of $7.9X10^{-5}$, or 1 in 12,500, for the probability of identity which assumes some relation. PIsibs is more relevant in this data set due to the fact that the Duluth population is a small population and the probability of two trees being related is a factor. Using GenoDive, we calculated genetic statistics (Table 4.) for just the Duluth population which included expected and observed levels of heterozygosity. H_0 and H_e are measures of heterozygosity in the population; a value of 0 represents a completely homozygous population (very little diversity), while a value of 1 represents a completely heterozygous

population (very high levels of diversity). Heterozygosity was high for all loci except GD15 (H_0 > 0.81).

Clonal Analysis

We found 12 matches from the Duluth population to the USDA-NPGS (Table 2). We Identified six Haralsons, and one of each Honeygold, Charlamoff, Oriole, Redwell, Redfree, and Jewett Red cultivars. Within Duluth we found 27 matches, representing 11 total genotypes, including the 6 Haralsons previously identified (Table 3). Note that all of these are matches within community samples and no matches were found to the UMD trial orchard. A total of 107 samples showed no matches, either to the USDA collection or to other samples in the Duluth area.

Discussion:

As a general convention, Peakall et al. (2006) suggests our PI is low enough to confirm clones if the inverse of population size is greater than PI. For our dataset, $N^{-1}=7.14X10^{-3}$ which is much greater than PIsibs($7.14X10^{-3}>7.9X10^{-5}$). Our PIsibs values reach this critical value with five loci (Fig. 1). For all loci except GD15 observed heterozygosity was higher than 0.81 (Table 4). This means that the Duluth population and the apple population as a whole (Gross et al. 2012; Cornille et al. 2012) are very diverse. Given the high values of H_0 and the low value of PIsibs we can confidently say any matched genotypes are truly clones and not just closely related individuals.

We formed two hypotheses to begin this experiment. We first hypothesized that apple trees found in the community would be the same varieties found in the UMD trial orchard. We found no supporting data for this hypothesis. Table 3 shows all clones we found within the Duluth population and none are found in the UMD trial orchard. The orchard was active from about 1910 to 1960, which means there is a possibility that clones were grafted from the orchard but died sometime in the last 50 years. Another possibility is that apple trees have been for sale from other sources for many years, allowing community members to 'import' their apples, effectively diluting the orchards signature in the population.

Our second hypothesis was that one species would dominate the population. We found that of our 140 samples, six were Haralson, three were unknown variety designated 1784 and nine more varieties were represented just twice in the data. We ran a χ^2 test against the null hypothesis of equal representation in the population, and found that our value was much less than the critical value ($12.1 \ll 26.3$). Although this data is not statistically significant, and so we fail to reject the

null, we can empirically say there were more Haralson varieties than any other variety present in the data set (Haralson represents 4.3% of tested apples). This non-significance could possibly be an artifact of small sample size. Upon further research, the Haralson apple was introduced by the Minnesota horticulture Research Center in 1923 (University of Minnesota Landscape Arboretum 2007), with the intent of introducing a cold hardy apple tree able to survive the harsh winters. With this in mind, it is easy to see why it would be at a high proportion here in Duluth.

Conclusion:

Apple trees are propagated by grafting, resulting in duplicate genetic profiles for both the parent and daughter tree. This unique feature can therefore be used to not only identify one cultivar from another, but also to identify clones within an unknown population. In this experiment we used nine SSRs to determine identity of Duluth community apple trees. We identified 12 apple trees, six of which were of the Haralson variety. Although we found that no community tree matched the UMD trial orchard trees, we did find many trees that matched within the Duluth community population, perhaps suggesting a local favorite. Finally, it is interesting to note the 107 tested trees that went unmatched to anything else. This number suggests that the Duluth area carries many unique varieties found nowhere else. More research will be done following this project with a larger data set and more in-depth analysis.

Bibliography

Cornille, A. et al,. (2012). New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics* 8:1-13.

Gross, B. L., Volk, G. M., Richards C. M, Forsline, P. L., Fazio, G., Chao C. T., (2012). Identification of “duplicate” accessions within the USDA-ARS National Plant Germplasm System *Malus* collection. *Journal of the American Society of Horticultural Science* 137:333-342.

Hale C. M., (2012). Personal communications

Hokanson, S. C., Szewc-McFadden, A. K., Lamboy, W. F., McFerson, J. R., (1998). Microsatellite (SSR) markers reveal genetic identities, genetic diversity and relationships in a *Malus × domestica* borkh. core subset collection. *Theoretical and Applied Genetics* 97:671-683.

Meirmans, P., Van Tienderen, P., (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* 4:792-794.

Peakal, R., Ebert, D., Cunningham, R., Lindenmayer, D., (2006). Mark-recapture by genetic tagging reveals restricted movements by bush rats (*Rattus fuscipes*) in a fragmented landscape. *Journal of Zoology* 268:207-216.

Schuelke, M. (2000). An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18:233-234.

University of Minnesota Landscape Arboretum Horticultural Research Center. 3675 Arboretum Drive, Chaska MN 55318. Horticultural Research Center. By Luby J. University of Minnesota, 28 Nov. 2007. Web. 09 Dec. 2013. <http://www.arboretum.umn.edu/bkgdresearchcenter.aspx>.

Tables and Figures

Table 1. PCR product multiplex. Groups were chosen based on separation of fragment sizes.

Multiplex Group	Primer Name	Size	Annealing Temp	Dye Color
Group 1	GD142	123-177	60	PET
	GD12	143-203	55	NED
	CH02d08	202-260	55	FAM
Group 2	GD147	117-177	55	FAM
	GD96	147-209	55	NED
	GD162	202-272	55	PET
Group 3	CH01h01	90-150	60	PET
	CH01f02	152-222	60	NED
	GD15	146-158	55	FAM

Table 2. Duluth Tree samples that matched to accessions in the USDA national collection. Matches were calculated with a threshold T=2.

Duluth Sample ID	USDA Cultivar ID	USDA collection Plant Identification number	Year of first record	Country of origin
HHO4	Haralson	589469	1913	United States
HH11	Haralson	589469	1913	United States
SJC_01	Haralson	589469	1913	United States
NAW_01	Haralson	589469	1913	United States
DAW_01	Haralson	589469	1913	United States
CGH_01	Haralson	589469	1913	United States
HHO8	Honeygold	588939	1935	United States
BD-66	Charlamoff	588978	1700	Former Soviet Union
N3	Oriole	589897	1914	United States
KMN_03	Redwell	589901	1911	United States
HNO9	Redfree	594111	Unknown	United States
AP-26	Jewett Red	644185	1842	United States

Table 3. Samples that matched within the Duluth population. Tentative cultivar names are given if a sample was submitted as a 'known' cultivar, otherwise the numerical clone ID generated by GenoDive is given.

Sample	Tentative cultivar	Ownership of sample
HHO4	Haralson	Hale Orchard
HH11	Haralson	Hale Orchard
SJC_01	Haralson	Community
NAW_01	Haralson	Community
DAW_01	Haralson	Community
CGH_01	Haralson	Community
HNO3	Vol. Crab	Hale Orchard
HNO8	Vol. Crab	Hale Orchard
N4	1755	Nelson Orchard
N5	1755	Nelson Orchard
SRP_03	1781	Community
RAO_01	1781	Community
BAP_08	1784	Community
LAT_01	1784	Community
SKL_01	1784	Community
EEW_02	1785	Community
RAO_03	1785	Community
TSL_01	1787	Community

ASC_01	1787	Community
TSL_02	1796	Community
ENP_02	1796	Community
JWD_01	1807	Community
SJR_03	1807	Community
ENP_04	1813	Community
ENP_05	1813	Community
GHF_01	1822	Community
EAL_01	1822	Community

Table 4. Genetic diversity. Number of alleles, Probability of Identity (PI), observed heterozygosity (H_0 and expected Heterzygosity under Hardy-Weinburg equilibrium (H_e).

Locus	Alleles	PI	H_0	H_e
GD12	9	0.05418	0.819	0.777
GD15	3	0.01777	0.114	0.11
GD96	17	0.02754	0.872	0.881
GD142	16	0.6146	0.908	0.847
GD147	12	0.02619	0.857	0.76
GD162	15	0.01580	0.936	0.812
CH01h01	12	0.01228	0.85	0.813
CH01f02	20	0.02091	0.963	0.909
CH02d08	17	0.02993	0.855	0.817

Figure 1. Probability of identity for increasing locus combinations. Loci are arranged in the same order as shown in Table 1.