J. Amer. Soc. Hort. Sci. 134(2):228–235. 2009.

Selection of Stratified Core Sets Representing Wild Apple (*Malus sieversii*)

Christopher M. Richards[1], Gayle M. Volk, Patrick A. Reeves, Ann A. Reilley, and Adam D. Henk
National Center for Genetic Resources Preservation, U.S. Department of Agriculture, Fort Collins, CO 80521

Philip L. Forsline
Plant Genetic Resources Unit, U.S. Department of Agriculture, Geneva, NY 14456-0462

Herb S. Aldwinckle
Department of Plant Pathology, Cornell University, Geneva, NY 14456-0462

Additional index words. simple sequence repeat, genetic diversity, Kazakhstan, genotype, phenotype

Abstract. **We estimate the minimum core size necessary to maximally represent a portion of the U.S. Department of Agriculture's National Plant Germplasm System apple (*Malus*) collection. We have identified a subset of *Malus sieversii* individuals that complements the previously published core subsets for two collection sites within Kazakhstan. We compared the size and composition of this complementary subset with a core set composed without restrictions. Because the genetic structure of this species has been previously determined, we were able to identify the origin of individuals within this core set with respect to their geographic location and genetic lineage. In addition, this core set is structured in a way that samples all of the major genetic lineages identified in this collection. The resulting panel of genotypes captures a broad range of phenotypic and molecular variation throughout Kazakhstan. These samples will provide a manageable entry point into the larger collection and will be critical in developing a long-term strategy for ex situ wild *Malus* conservation.**

A central tenet of gene bank management is to make a collection useful. The applied value of these collections for crop improvement or gene discovery often depends upon fostering efficient utilization. The concept of core collections (or core sets) was initially proposed as a way to define, as a representative subset, the genetic diversity of a crop species (Brown, 1989a; Frankel, 1984). Thus, core collections provide an efficient entry point to the whole collection that is composed of a subset of diversity for researchers, breeders, and trait specialists. As a management tool, core collections have been proposed to capture the common and rare alleles within a fraction (5–10%) of the original collection (Brown, 1989b). Traditionally, core collections have been determined based on geographical and phenotypic characteristics (Crossa et al., 1993), but increasingly, genetic data has also been used to make selections (Liu et al., 2003; Marita et al., 2000; Ronfort et al., 2006). In many cases, a large collection may have developed targeted subsamples that are focused on specific traits or localities of interest (e.g., Ma et al., 2006).

It has been suggested that having core sets available for clonal collections is particularly desirable for vegetatively propagated clonal collections (Hodgkin et al., 1995). This germplasm is more expensive to maintain than orthodox seed collections because individuals are kept under field, greenhouse, or in vitro conditions rather than in long-term reduced temperature storage. Clones maintained in limited field plantings are subject to attrition through disease or bad weather. Such core sets may be more frequently requested and gene bank managers can plan on having appropriate propagules available when needed. Increased distribution of core sets can result in additional characterization data, thus increasing the utility of the larger collection the core set represents (Rubenstein et al., 2006; van Hintum, 1999). Seed-based core sets may be particularly useful in wild relative collections of clonally propagated crops. In these accessions, the objective is frequently to capture the allelic variation within the accession, but not necessarily any particular genotypes. These core sets can be used in a crossing design to preserve allelic variation segregating within populations of seeds, which can then be stored for longer periods of time at far lower cost (Volk et al., 2005).

U.S. Department of Agriculture (USDA) plant exploration teams collected *Malus sieversii* seeds and clones from Kazakhstan between 1989 and 1996 (Dzhangaliev, 2003; Forsline et al., 2003; Hokanson et al., 1997; Luby et al., 2001). Over one thousand trees derived from seeds collected during these trips have been planted and characterized at the USDA-Agricultural Research Service Plant Genetic Resources Unit (PGRU) in Geneva, NY. Samples for these trees representing eight collection sites in Kazakhstan and one collection site in Kyrgyzstan have been genotyped to determine the population structure of wild *M. sieversii* using seven highly variable microsatellite loci (Richards et al., 2009). Results from this analysis using standard population genetic approaches and Bayesian assignment methods identified four genetically distinct, stable clusters of individuals (Richards et al., 2009). Importantly, these clusters revealed a pattern of variation that

Received for publication 4 Nov. 2008. Accepted for publication 29 Dec. 2008.
Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.
[1]Corresponding author. E-mail: Chris.Richards@ars.usda.gov.

was not primarily defined among collection sites but rather among broad geographic regions. This regional pattern of differentiation revealed ongoing admixture that obscured site-specific differentiation.

Remarkable progress has been made in the generation of genotypic data in many agricultural taxa and the development of algorithms and bioinformatic tools used to guide the construction of core subsets. Many methods rely on initial stratification of the samples into groups that reflect some ecogeographic attribute or quantitative trait value (Brown, 1989a, 1995; Franco et al., 2005, 2006; Li et al., 2004). Stratification ensures that sampling is distributed among relevant groups defined beforehand. Alternatively, maximization strategies attempt to reduce redundancy in a core set without a priori stratification (Schoen and Brown, 1995). Maximization strategies have been developed that can be used to efficiently assemble core subsets based on character states such as alleles at molecular loci or values of quantitative traits (Brown, 1989b; Gouesnard et al., 2001; McKhann et al., 2004; Schoen and Brown, 1993, 1995). A key feature of this approach is that redundancy in the collection can be empirically assessed and the size of an appropriate core set can be established (Gouesnard et al., 2001).

While the collection itself represents an important source of variation for breeding improvement in *Malus*, the living orchard collection is at risk of continued seasonal mortality. Core subsets of individuals representing the genotypic and phenotypic diversity of two of the largest collection sites in Kazakhstan have been proposed in part to stem these losses by developing a long-term seed-based backup and to increase utility of this material by breeders and researchers worldwide (Volk et al., 2005). In addition, vegetative buds from the individuals in the core subset will be cryopreserved to ensure long-term availability. The two collection sites were considered separately to maintain putative site-specific environmental adaptations to drought and cold temperatures. Establishing a complementary core set among the other seven collection sites is the next step in developing a comprehensive conservation strategy for the entire Kazakhstan collection at PGRU.

A key feature of determining composition of the proposed core set in this study is the availability of an estimate of the genetic structure of this collection (Richards et al., 2009). This study used population genetics of diversity and linkage disequilibrium between alleles at the sampled loci as a metric used to partition the genotypes into groups that share common ancestry. One advantage of this method is that admixture among geographic regions can be detected and quantified.

In this article, we estimated the minimum core size necessary to maximally represent the diversity of the *M. sieversii* collection. Our objective was to identify a subset of *M. sieversii* individuals that complements the existing core subsets for two collection sites within Kazakhstan. This study investigated the size and composition of this complementary subset. We were particularly interested in the origin of individuals in these selected sets—what sites and what genetic lineages are represented? Inclusion of diverse sites and genetic lineages in core collections is key to ensure that full representation is achieved.

Materials and Methods

Collection materials and DNA extraction. *Malus sieversii* seeds were collected from wild trees during plant explorations to Kazakhstan in 1989, 1993, 1995, and 1996. Clones of individuals collected in Kazakhstan and classified as elite due to unusual or desirable characteristics were not included in these analyses. DNA was extracted from duplicate leaf samples from 961 seedling accessions available in the field collections in 2003. The individual composition of the core subsets identified by Volk et al. (2005) was modified slightly. Specifically, seven accessions were added to core sets 6 and 9. These additions were necessary to offset variation in vigor among the original selections and they brought the total number of accessions within the two site-specific core collections to 77. Seven amplified microsatellite loci yielding 103 alleles were separated on a gel-based system (LI-COR, Lincoln, NE) as previously described (Richards et al., 2009; Volk et al., 2005). The simple sequence repeats (SSR) were amplified using unlinked primers GD12, GD15, GD96, GD100, GD142, GD147, and GD162 (Hemmat et al., 2003; Hokanson et al., 2001). Phenotypic data from 21 continuous traits were categorically classified according to standards described in the publicly available USDA Germplasm Resources Information Network database (USDA, 2004).

Data analysis. In 2007, 797 of the 961 seeding trees were alive and available in the field collection. Genotypic and phenotypic data for these 797 accessions were considered for the construction of core collections. All analyses used the maximization algorithm in the software package MSTRAT (Gouesnard et al., 2001), based upon the maximization strategy proposed by Schoen and Brown (1995). Briefly, this method treats each allele and each quantitative trait category as a unique character state. The object is to identify the smallest subset of individuals that contains all the character states—a set that is maximized for character state variation (Gouesnard et al., 2001). For some analyses, MSTRAT was made to include the previously identified genotypes used in core sets developed for two of the nine collection sites genotyped (Volk et al., 2005). In these cases, maximization focused on complementing variation not included in the original cores by identifying novel variation in the other seven sites.

As a benchmark, we estimated the size of a collection needed to capture about 90% to 95% of the total variation by using a feature of MSTRAT that measures the fraction of total diversity obtained in core subsets of varying size. If each genotype in the larger collection contributed some unique character, there would be a linear relationship between variation captured and sample size. However, variation is commonly structured, especially in natural populations where dispersal limits panmixia (Hamrick and Godt, 1997). In these cases, the fraction of total diversity (measured in character states) of a sample plateaus at a certain size, similar to a saturation curve where there is a diminishing return on diversity after a certain sample size is reached. These ''redundancy'' curves were developed using the mean fraction of diversity captured in five independent sampling runs. The inflection point of the resulting curvilinear plot can be used to find the optimal core sample size (Gouesnard et al., 2001). We estimated the optimal core size in the original set of 961 genotypes and the sample of 797 genotypes that were healthy and flowering in 2007. In addition, we compared the optimal core size among these datasets when using molecular data (7 loci, 103 alleles), phenotypic data (21 traits, 114 total trait states), or both. The quantitative metric used for this comparison was the fraction of total character states retained in the core set. The resulting diversity of these

cores assembled using maximization was compared with core sets assembled at random. The difference between the two sampling curves illustrates the net gain in diversity realized through maximization, and provides a relative measure of core collection success in capturing representative variation. Once an appropriate core size was identified, the composition of this subset was examined. In many instances, there were several equally diverse core sets. To develop a consensus set of genotypes, we examined 10 possible core sets for each dataset. We chose the set that contained the most commonly found genotypes among the 10 replicate core sets.

We examined the distribution of collection sites and genetic clusters represented within a core subset from the complete set of 961 genotypes (core-SSR), or a core subset using the genotypic and phenotypic data for the 797 individuals living in 2007.

Results

Redundancy curves in the full set of 961 genotypes show that regardless of the source of the data (molecular or phenotypic), cores maximized for character diversity always capture more diversity than a similarly sized, randomly assembled core (Fig. 1). However, data types capture diversity at different levels of efficiency. Phenotypic data saturated earlier—it took as few as 27 individuals to capture 95% of the phenotypic diversity. While the number of states was high for these quantitative traits, saturation required few individuals. This is most likely because many of the agricultural traits showed high covariance. In contrast, it took 84 individuals to capture 91% of the genotypic diversity [subsequently referred to as core-SSR, n = 84 (Fig. 1)].

Redundancy curves were also developed for a set of genotypes to complement the diversity in the established cores for sites 6 and 9 (Fig. 2). For these data, we considered only the 797 healthy individuals. The 77 individuals representing the site 6 and 9 core collections (Volk et al., 2005) were indexed in a way that they became a mandatory part of the resulting core. A complementary third core of 35 individuals captured 94% of the measured genetic and phenotypic diversity of the entire 797 seedling dataset (Table 1). In contrast, when individuals were randomly selected from the population of 797, 445 individuals were required to capture a comparable level of diversity (Fig. 1). These selected individuals (complementary core, n = 35) exhibit desirable characteristics such as disease resistance and fruit quality traits. For example, 54% of the individuals in the complementary core are resistant to fire blight (*Erwinia amylovora*) and 34% are resistant to apple scab (*Venturia inaequalis*).

The contribution of each collection site to each core set is shown in Fig. 2. The histogram shows the proportion of genotypes in a core set drawn from each collection site. The data confirm that when core subsets from sites 6 and site 9 (totaling 77 individuals) are forced to be included into a core of the entire collection, the additional 35 individuals (complementary core, n = 35) that are needed to capture the remaining variation are chosen primarily from sites 3, 5, 7, 11, and 12 (Fig. 2). The core-SSR (n = 84) developed from all the available 965 genotypes (Table 2) included individuals drawn roughly in proportion to the number of samples collected at each site. The one exception was in site 5, which contributed disproportionally to the core set, most likely due to the presence of rare private alleles.

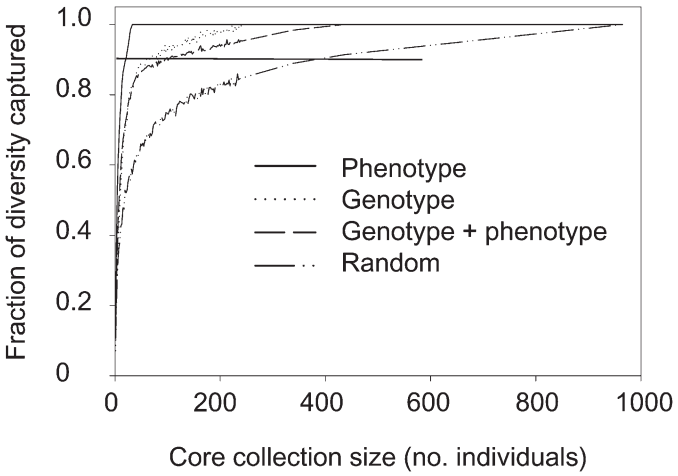

Fig. 1. Diversity redundancy curves for all 961 *Malus sieversii* individuals in the complete data set. Plots compare the amount of diversity retained in cores maximized for trait diversity based on available phenotypic, genotypic, or phenotypic and genotypic data (top three curves) and similarly sized, randomly assembled cores (bottom curve). To capture phenotypic trait variation, fewer individuals could be used than were necessary to capture an equal percentage of the genotypic variation.

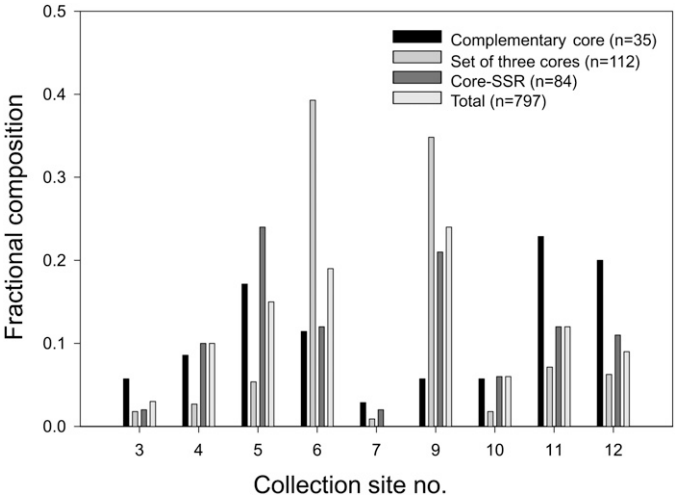

Fig. 2. Fractional composition of *Malus sieversii* core collections among the nine collection sites in Kazakhstan and Kyrgyzstan. Each core set was developed with different source data and objectives. The total (n = 797) represents the distribution of samples that were healthy and flowering in 2007. Core-SSR (n = 84) represents the subset of individuals selected using SSR data from the total 961 individuals genotyped. Set of three cores (n = 112) represents a collection where the previous site 6 and 9 core sets were locked in (77 genotypes) and 35 additional genotypes were selected to capture the most diversity in the total collection. Complementary new core (n = 35) represents the genotypes identified in this study that complement the previous two site specific core sets.

The contribution of each genetic cluster to each core set is shown in Fig. 3. The histogram shows the proportion of genotypes in each core set that were selected from each cluster. Comparison of each core set shows some slight differences in composition, especially where the core set composition differs from the total (n = 797). These discrepancies may be due to

Table 1. *Malus sieversii* phenotypic characterization data for the new complementary core of 35 individuals.

Identification no.	Site no.	Cluster no.	Yr collected	Characterization year	Fire blight resistance	Scab resistance	Flesh color	Flesh firmness	Flesh flavor	Flesh oxidation (%)	Fruit ground color[y]
GMAL 3541.l[z]	5	1	1993	2001	Resistant	Resistant	Cream, green	Semifirm	Astringent	>10	Lt. green
GMAL 3583.c[z]	7	4	1993	—	—	—	—	—	—	—	—
GMAL 3616.j	9	1	1995	2004	Susceptible	Susceptible	White	Soft	Subacid	5–10	Green
GMAL 3627.h	9	1	1995	—	Susceptible	Resistant	—	—	—	—	—
GMAL 3688.h	6	2	1995	2002	Resistant	Susceptible	Cream, green	Firm	Astringent	>10	Lt. green
GMAL 3689.i	6	1	1995	2000	Resistant	Resistant	Green	Firm	Astringent	5–10	Lt. Yellow
GMAL 3691.b	6	2	1995	2002	Resistant	Susceptible	Cream	Semifirm	Subacid	>10	Lt. green
GMAL 4002.g	6	2	1995	2004	Susceptible	Susceptible	Cream, green	Hard	Acid	1–4	Green
GMAL 4011.p[z]	10	1	1996	—	Susceptible	Susceptible	—	—	—	—	—
GMAL 4028.g	4	1	1996	—	Susceptible	Susceptible	—	—	—	—	—
GMAL 4032.s	5	1	1996	2002	Resistant	Susceptible	Cream, green	Semifirm	Subacid	>10	Lt. green
GMAL 4036.o	5	1	1996	—	Susceptible	Susceptible	—	—	—	—	—
GMAL 4039.x	5	1	1996	2003	Susceptible	Resistant	Cream	Semifirm	Subacid	5–10	Green yellow
GMAL 4051.v	5	4	1996	2003	Susceptible	Susceptible	White	Semifirm	Subacid	>10	Lt. green
GMAL 4053.d[z]	11	3	1996	—	Resistant	Susceptible	—	—	—	—	—
GMAL 4053.t	11	3	1996	2005	Susceptible	Susceptible	White, green	Hard	Subacid	>10	Lt. green
GMAL 4053.v	11	3	1996	2004	Susceptible	Susceptible	Yellow	Firm	Sweet	0–1	Green
GMAL 4054.a	12	3	1996	2005	Susceptible	Susceptible	Cream	Hard	Astringent	>10	Lt. green
GMAL 4054.e	12	3	1996	—	Susceptible	Susceptible	—	—	—	—	—
GMAL 4056.d	12	3	1996	—	Susceptible	Resistant	—	—	—	—	—
GMAL 4076.d[z]	10	2	1996	2005	Resistant	Susceptible	Green	Firm	Acid	>10	Lt. Green
GMAL 4082.b	3	1	1996	—	Resistant	Resistant	—	—	—	—	—
GMAL 4087.b[z]	3	2	1996	—	Susceptible	Resistant	—	—	—	—	—
GMAL 4206.f	4	4	1996	2004	Resistant	Susceptible	White	Semifirm	Acid	>10	Green, lt. yellow
GMAL 4209.c	4	1	1996	—	Resistant	Resistant	—	—	—	—	—
GMAL 4237.a[z]	5	1	1996	—	Resistant	Resistant	—	—	—	—	—
GMAL 4237.d[z]	5	1	1996	—	Resistant	Susceptible	—	—	—	—	—
GMAL 4278.a[z]	11	3	1996	—	Resistant	Susceptible	—	—	—	—	—
GMAL 4296.d	11	2	1996	2004	Resistant	Susceptible	White	Semifirm	Subacid	1–4	Red
GMAL 4304.b	11	1	1996	2005	Resistant	Resistant	Cream	Firm	Sweet	5–10	Yellow
GMAL 4304.e	11	1	1996	2002	Resistant	Resistant	Cream, yellow	Soft	Sweet	>10	Lt. green
GMAL 4309.b	12	4	1996	—	Susceptible	Susceptible	—	—	—	—	—
GMAL 4309.c	12	4	1996	2002	Resistant	Susceptible	Cream, green	Semifirm	Aromatic	>10	Green
GMAL 4309.d	12	4	1996	2004	Resistant	Susceptible	Cream	Firm	Subacid	>10	Lt. green
GMAL 4311.a	12	2	1996	—	Resistant	Resistant	—	—	—	—	—

[z]Accessions included in core-SSR set (n = 84).
[y]Fruit ground color was classified as: green, light (lt.) green, green yellow, light yellow, yellow, or red.
[x]Fruit juiciness was determined based on apple weight/apple volume (specific gravity) of the mean of five apples at maturity: very dry (<0.75), dry (0.76–0.80), medium (0.81–0.85), or moderately (mod.) juicy (0.86–0.90).
[w]Fruit russet type was classified as extremely (extr.) fine, medium heavy (somewhat rough), and surface cracked.
[v]Harvest season was classified with reference to 'Delicious': extremely early (>60 d before 'Delicious'), very early (50–60 d before 'Delicious'), early (30–50 d before 'Delicious'), medium-early (20–30 d before 'Delicious'), medium (same time as 'Delicious'), or very late (20–30 d later than Delicious').

continued on next page

properties of the core set criteria (such as forcing the inclusion of 77 genotypes) or to diversity of the genetic cluster. The set of three cores (n-112) are composed of genotypes in proportion to the size of each of the clusters, whereas core-SSR (n = 84) draws more heavily on clusters 3 and 4. The new proposed complementary core (n = 35) has a higher representation of individuals selected from clusters 3 and 4 (20% and 14%, respectively) than would be predicted by the size of the cluster (total n = 797) (Fig. 3). Thus, the proposed new core is heavily represented by individuals drawn from the smaller genetic clusters 3 and 4.

Discussion

The *M. sieversii* seedling collection maintained in the field in Geneva, NY, has over 1000 inventories that represent 108 mother trees. This collection displays high levels of diversity at the phenotypic and genotypic level, but its size makes it unwieldy for many research and breeding efforts. The *M. sieversii* collection becomes more manageable when core sets of individuals are available that capture this diversity at the phenotypic and genotypic level. The phenotypic traits included in these analyses include disease resistance, quality, and yield characteristics, all of which are important considerations in breeding programs. The inclusion of data collected from unlinked genotypic markers makes the proposed core sets potentially more diverse than they would have been if only phenotypic data were considered.

The complementary core (n = 35) increases the representation of genotypes among collection sites and genetic lineages that were not represented in the core sets proposed for sites 6 and 9 (Volk et al., 2005). The development of three independent complementary cores for *M. sieversii* provides researchers with

Table 1. Continued.

Fruit juiciness[x]	Fruit wt (g)	Fruit overcolor	Overcolor on fruit (%)	Fruit overcolor pattern	Fruit surface with russet (%)	Fruit russet location	Fruit russet type[w]	Fruit shape	Fruit shape uniformity	Fruit size uniformity	Fruit texture	Harvest season[v]	Soluble solids (%)
Moderate	<50	Pink	25	Striped	0	—	—	Globose	Uniform	Uniform	Coarse	Very early	11.3
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dry	100–150	Pink	10	Striped	1	Pedicel	Extr. fine	Globose	Uniform	Variable	Medium	Medium	12.2
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dry	<50	None	—	—	4	Pedicel, calyx	Extr. fine	Flat-globose	Uniform	Uniform	Coarse	Early	11.1
Moderate	<50	Yellow	30	Blush	—	—	—	Flat-standard	Uniform	Uniform	Fine	Extremely early	10.1
Dry	<50	—	—	—	0	—	—	Globose	Uniform	Uniform	Medium	Very early	11
Dry	50–100	Pink	15	Blush	60	Pedicel, calyx	Surface cracked	Globose	Uniform	Variable	Medium	—	12.9
—	—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dry	50–100	Red	40	Striped	2	Pedicel	Extr. fine	Globose-conical	Variable	Variable	Medium	Extremely early	10.10
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dry	50–100	Red	80	Striped	1	Pedicel	Extr. fine	Globose-conical	Uniform	Uniform	Medium	Medium	12
Dry	50–100	None	0	None	12	Pedicel	Medium heavy	Conical	Uniform	Uniform	Medium	Medium	10.8
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Moderate	50–100	Pink	25	Striped	25	Pedicel	Surface cracked	Conical	Variable	Variable	Fine	—	12.5
Moderate	50–100	Pink	25	Blush	8	Pedicel	Medium	Globose	Uniform	Variable	Fine	—	13.1
Very dry	50–100	Red orange	40	Blush	20	Pedicel	Extr. fine	Conical	Uniform	Uniform	Fine	—	17
—	—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dry	<50	Red yellow	50	Striped	10	Pedicel	Surface cracked	Globose	Uniform	Uniform	Fine	—	12.8
—	—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dry	<50	Yellow	60	Blush	0	—	—	Globose	Uniform	Uniform	Fine	—	11.3
—	—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dry	<50	Dark red	70	Striped	4	Pedicel	Extr. fine	Oblong	Variable	Variable	Medium	—	12
Mod. juicy	<50	None	0	—	2	Pedicel	Extr. fine	Flat-globose	Variable	Variable	Fine	—	14.4
Medium	<50	Red	75	Striped	1	Pedicel, calyx	Extr. fine	Oblong	Variable	Variable	Medium	Medium-early	14.2
—	—	—	—	—	—	—	—	—	—	—	—	—	—
Mod. juicy	<50	Pink	20	Blush	70	Entire	Surface cracked	Globose	Uniform	Uniform	Coarse	Very late	14.8
Mod. juicy	50–100	Pink	35	Blush	5	Pedicel, calyx	Extr. fine	Globose-conical	Uniform	Uniform	Medium	—	15.5
—	—	—	—	—	—	—	—	—	—	—	—	—	—

tools that allow them to select the group of individuals that are most relevant to their research goals. Spatial genetic patterns specific to sites 6 and 9 can be evaluated in site-specific cores, and the complementary core of 35 serves to capture the diversity that was not available at those locations. Researchers who are interested in trees that may be particularly drought tolerant or cold hardy can select the core collections targeted to sites 6 and 9, respectively. Alternatively, those interested in evaluating a representative subset of the USDA *M. sieversii* collection can choose to use the combined core set of 112 individuals.

The use of maximization algorithms to identify core subsets has resulted in cores that capture allelic, geographic, and phenotypic diversity (Balfourier et al., 2007). New algorithms continue to be proposed that may also capture diversity by measuring distances between accessions within defined groups (Jansen and van Hintum, 2007), by least distance stepwise sampling (Wang et al., 2007), or by sampling a single individual from sets of clusters (Franco et al., 2005). The quantitative metrics used to assess the efficiency of these algorithms often describe how the mean and variance of a trait value within the core compares with the larger collection (e.g., Upadhyaya and Ortiz, 2001). We selected the maximization method because we knew that the geographic boundaries of genetic variation were more diffuse in this species due to ongoing admixture.

Implicit in these core collections is the assumption that core sets maximized for diversity using a set of specific attributes (molecular or phenotypic) are in fact representative of diversity elsewhere in the genomes of the selected individuals (Bataillon et al., 1996). Validation of this assumption comes from assessing the retention of variation at independent loci in the core (Le Cunff et al., 2008; McKhann et al., 2004; Ronfort et al., 2006). Evidence from simulation analysis suggests that these validation approaches will support core set selection more often in inbreeding species (Bataillon et al., 1996). While we do not use independent loci to validate these selections, we show that the core set composition reflects not only geographic diversity but also, and most importantly, the genetic diversity at the level of lineages. In studies of natural systems, a priori designations of the units that comprise populations or clusters are often based upon geographical criteria such as the collection site where ecological and environmental conditions can be assessed. Increasingly, studies of structure rely on novel model-based clustering methods that use a Bayesian analytical procedure to simultaneously reveal cryptic population structure and assign individuals to clusters (Huelsenbeck and Andolfatto, 2007;

Table 2. Core set of 84 *Malus sieversii* individuals (core-SSR) identified using genotypic data. Individuals are classified according to collection site, family (arbitrary identification number), and cluster.

Identification no.	Site no.	Family no.	Cluster no.	Identification no.	Site no.	Family no.	Cluster no.
GMAL 3541.l[z]	5	6	1	GMAL 4038.n	5	63	1
GMAL 3544.j	5	7	1	GMAL 4038.t	5	63	1
GMAL 3552.v[z]	5	10	1	GMAL 4039.d	5	64	1
GMAL 3574.a[z]	7	14	1	GMAL 4039.s	5	64	1
GMAL 3583.c	7	15	4	GMAL 4039.t	5	64	1
GMAL 3607.h	9	16	1	GMAL 4039.v	5	64	1
GMAL 3607.k	9	16	1	GMAL 4039.w	5	64	1
GMAL 3610.a	9	18	1	GMAL 4047.n	11	65	2
GMAL 3616.b	9	20	1	GMAL 4047.s	11	65	2
GMAL 3623.f	9	24	1	GMAL 4049.m	11	66	3
GMAL 3629.g	9	28	1	GMAL 4049.w	11	66	3
GMAL 3631.m	9	29	1	GMAL 4051.f	11	67	4
GMAL 3635.i	9	30	1	GMAL 4051.o	11	67	4
GMAL 3636.h	9	31	3	GMAL 4053.d[z]	11	68	3
GMAL 3637.d	9	32	1	GMAL 4054.d	12	69	3
GMAL 3638.c	9	33	1	GMAL 4054.f	12	69	3
GMAL 3638.j	9	33	1	GMAL 4054.m	12	69	3
GMAL 3682.d	6	35	1	GMAL 4054.aa	12	69	3
GMAL 3682.i	6	35	2	GMAL 4056.a	12	71	1
GMAL 3683.k	6	36	2	GMAL 4056.o	12	71	4
GMAL 3687.a	6	39	2	GMAL 4056.p	12	71	3
GMAL 3687.h	6	39	2	GMAL 4056.q	12	71	3
GMAL 3688.c	6	40	2	GMAL 4059.a	10	72	2
GMAL 3689.a	6	41	2	GMAL 4059.f	10	72	2
GMAL 3689.g	6	41	1	GMAL 4068.b	10	74	2
GMAL 3762.h	9	44	1	GMAL 4076.d[z]	10	76	2
GMAL 3781.c	9	47	1	GMAL 4082.d	3	77	2
GMAL 3781.f	9	47	1	GMAL 4087.b	3	79	2
GMAL 3784.g	9	48	1	GMAL 4155.b	9	82	1
GMAL 3975.d	6	50	2	GMAL 4177.e	4	85	1
GMAL 3989.a	6	51	2	GMAL 4179.b	4	87	1
GMAL 4011.p[z]	10	55	1	GMAL 4179.f	4	87	1
GMAL 4020.f	9	56	1	GMAL 4179.g	4	87	1
GMAL 4028.h	4	58	2	GMAL 4198.e	4	89	1
GMAL 4028.v	4	58	1	GMAL 4211.g	4	93	1
GMAL 4032.i	5	59	1	GMAL 4237.a[z]	5	95	1
GMAL 4032.r	5	59	1	GMAL 4237.d[z]	5	95	1
GMAL 4032.t	5	59	1	GMAL 4238.c	5	96	1
GMAL 4032.w	5	59	1	GMAL 4278.a[z]	11	99	3
GMAL 4036.l	5	61	1	GMAL 4290.g	11	101	2
GMAL 4036.m	5	61	1	GMAL 4296.e	11	103	2
GMAL 4038.j	5	63	2	GMAL 4312.d	12	107	2

[z]Accessions included in the complementary core (n = 35) set.

Pritchard et al., 2000). The proposed core set specifically targets defined genetic clusters that represent different ancestry within the natural populations. Collections based on diverse genotypic data may have superior representativeness than those based on phenotypic data (Hu et al., 2000). This may be especially true in wild germplasm collections where phenotypic similarity may mask substantial genotypic diversity (Tanksley and McCouch, 1997).

The three distinct core collections of 40 (site 6), 37 (site 9), and 35 (complementary core, n = 35) individuals capture over 90% of the total diversity in the larger collection. The core sets include ≈14% of the individuals in the PGRU *M. sieversii* field collection. The trees included in this set of 112 will be repropagated and maintained indefinitely as clones in the main *Malus* collection. Additional data will be collected for accessions that have not yet been thoroughly phenotyped. The trees in the site 6 and site 9 core sets have been included in a large-scale hand-pollination crossing effort to generate sets of seeds that represent the genotypes of each core sets. Genotyping efforts are underway to confirm that these sets are indeed representative of the core diversity. In Spring 2008, the trees in the new core of 35 *M. sieversii* individuals will be crossed in a similar manner to produce seed lots that represent the diversity of this core set. The seed lots for each of the three *M. sieversii*

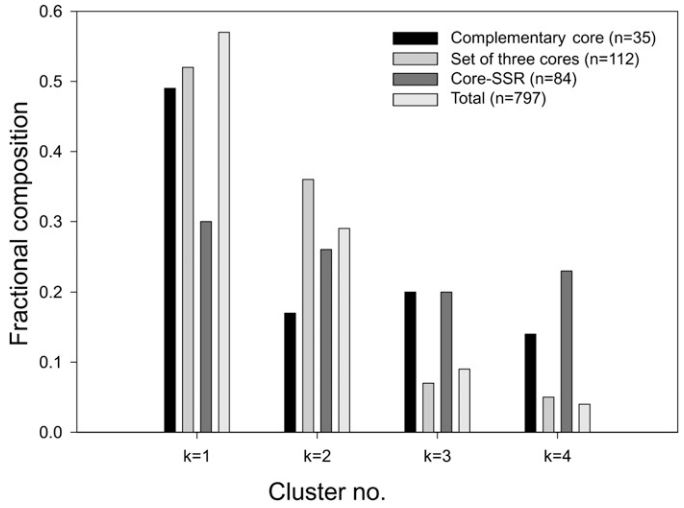

Fig. 3. Fractional composition of *Malus sieversii* core collections among four genetic clusters within the collection of viable genotypes from Kazakhstan and Kyrgyzstan. The histogram compares the numerical distribution of samples in core sets among four genetic lineages (clusters) found in the 961 *M. sieversii* individuals sampled. The total (n = 797) represents the distribution of samples that were healthy and flowering in 2007. Core-SSR (n = 84) represents the subset of individuals selected using SSR data from the total 961 individuals genotyped. Set of three cores (n = 112) represents a collection where the previous site 6 and 9 core sets were locked in (77 genotypes) and 35 additional genotypes were selected to capture the most diversity in the total collection. Complementary new core (n = 35) represents the genotypes identified in this study that complement the previous two site specific core sets.

core sets will be made available for distribution for research purposes.

Literature Cited

Balfourier, F., V. Roussel, P. Strelchenko, F. Exbrayat-Vinson, P. Sourdille, G. Boutet, J. Koenig, C. Ravel, O. Mitrofanova, M. Michel Beckert, and G. Charmet. 2007. A worldwide bread wheat core collection arrayed in a 384-well plate. Theor. Appl. Genet. 114:1265–1275.

Bataillon, T.M., J.L. David, and D.J. Schoen. 1996. Neutral genetic markers and conservation genetics: Simulated germplasm collections. Genetics 14:409–417.

Brown, A.H.D. 1989a. The case for core collections, p. 136–156. In: A.H.D. Brown, O. Frankel, D.R. Marshall, and J.T. Williams (eds.). The use of plant genetic resources. Cambridge University Press, Cambridge, UK.

Brown, A.H.D. 1989b. Core collections: A practical approach to genetic resources management. Genome 31:818–824.

Brown, A.H.D. 1995. The core collection at the crossroads, p. 3–19. In: T. Hodgkin, A.H.D. Brown, T.J.L. van Hintum, and E.A.V. Morales (eds.). Core collections of plant genetic resources. Wiley, Chichester, UK.

Crossa, J., C.M. Hernandez, P. Bretting, S.A. Eberhart, and S. Taba. 1993. Statistical genetic considerations for maintaining germplasm collections. Theor. Appl. Genet. 86:673–678.

Dzhangaliev, A.D. 2003. The wild apple tree of Kazakhstan. Hort. Rev. (Amer. Soc. Hort. Sci.) 29:63–303.

Forsline, P.L., H.S. Aldwinckle, E.E. Dickson, and S.C. Hokanson. 2003. Collection, maintenance, characterization, and utilization of wild apples from central Asia. Hort. Rev. (Amer. Soc. Hort. Sci.) 29:1–61.

Franco, J., J. Crossa, M. Warburton, and S. Taba. 2006. Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci. 46:854–864.

Franco, J., J. Crossa, S. Taba, and H. Shands. 2005. A sampling strategy for conserving genetic diversity when forming core subsets. Crop Sci. 45:1035–1044.

Frankel, O.H. 1984. Genetic perspectives on germplasm conservation, p. 161–170. In: W. Arber, K. Llimensee, W.L. Peacock, and P. Starlinger (eds.). Genetic manipulation: Impact on man and society. Cambridge University Press, Cambridge, UK.

Gouesnard, B., T.M. Bataillon, G. Decoux, C. Rozale, D.J. Schoen, and J.L. David. 2001. MSTRAT: An algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J. Hered. 92:93–94.

Hamrick, J.L. and M.J.W. Godt. 1997. Allozyme diversity in cultivated crops. Crop Sci. 37:26–30.

Hemmat, M., N.F. Weeden, and S.K. Brown. 2003. Mapping and evaluation of *Malus* ×*domestica* microsatellites in apple and pear. J. Amer. Soc. Hort. Sci. 128:515–520.

Hodgkin, T., A.H.D. Brown, T.J.L. van Hintum, and E.A.V. Morales. 1995. Future directions, p. 253–259. In: T. Hodgkin, A.H.D. Brown, T. van Hintum, and E.A.V. Morales (eds.). Core collections of plant genetic resources. Intl. Plant Genet. Resources Inst./Wiley-Sayce, Rome.

Hokanson, S.C., J.R. McFerson, P.L. Forsline, W.F. Lamboy, A.D. Djangaliev, and H.S. Aldwinckle. 1997. Collecting and managing wild *Malus* germplasm in its center of diversity. HortScience 32:173–176.

Hokanson, S.C., W.F. Lamboy, A.K. Szewc-McFadden, and J.R. McFerson. 2001. Microsatellite (SSR) variation in a collection of *Malus* (apple) species and hybrids. Euphytica 118:281–294.

Hu, J., J. Zhu, and H.M. Xu. 2000. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. Theor. Appl. Genet. 101:264–268.

Huelsenbeck, J.P. and P. Andolfatto. 2007. Inference of population structure under a Dirichlet process model. Genetics 175:1787–1802.

Jansen, J. and T. van Hintum. 2007. Genetic distance sampling: A novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. Theor. Appl. Genet. 114:421–428.

Le Cunff, L., A. Fournier-Level, V. Laucou, S. Vezzulli, T. Lacombe, A.F. Adam-Blondon, J.M. Boursiquot, and P. This. 2008. Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. *sativa*. BMC Plant Biol. 8:31.

Li, C.T., C.H. Shi, J.G. Wu, H.M. Xu, H.Z. Zhang, and Y.L. Ren. 2004. Methods of developing core collections based on the predicted genotypic value of rice (*Oryza sativa* L.). Theor. Appl. Genet. 108:1172–1176.

Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128.

Luby, J., P. Forsline, H. Aldwinckle, V. Bus, and M. Giebel. 2001. Silk road apples: Collection, evaluation, and utilization of *Malus sieversii* L. from central Asia. HortScience 36:225–231.

Ma, Y.S., W.H. Wang, L.X. Wang, F.M. Ma, R.Z. Chang, and L.J. Qiu. 2006. Genetic diversity of soybean and establishment of a core collection focused on resistance to soybean cyst nematode. J. Integr. Plant Biol. 48:722–731.

Marita, J.M., J.M. Rodriguez, and J. Nienhuis. 2000. Development of an algorithm identifying maximally diverse core collections. Genet. Resources Crop Evol. 47:515–526.

McKhann, H.I., C. Camilleri, A. Berard, T. Bataillon, J.L. David, X. Reboud, V. Le Corre, C. Caloustian, I.G. Gut, and D. Brunel. 2004. Nested core collections maximizing genetic diversity in *Arabidopsis thaliana.*. Plant J. 38:193–202.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Richards, C.M., G.M. Volk, A.A. Reilley, A.D. Henk, D.R. Lockwood, P.A. Reeves, and P.L. Forsline. 2009. Genetic diversity and population structure in *Malus sieversii*, a wild progenitor species of domesticated apple. Tree Genet. Genomes, doi: 10.1007/s11295-008-0190-9.

Ronfort, J., T. Bataillon, S. Santoni, M. Delalande, J.L. David, and J.M. Prosperi. 2006. Microsatellite diversity and broad scale geographic structure in a model legume: Building a set of nested core collection for studying naturally occurring variation in *Medicago truncatula*. BMC Plant Biol. 6(28), doi: 10.1186/1471-2229-6-28.

Rubenstein, K.D., M. Smale, and M.P. Widrlechner. 2006. Demand for genetic resources and the U.S. National Plant Germplasm System. Crop Sci. 46:1021–1031.

Schoen, D.J. and A.H.D. Brown. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. USA 90:10623–10627.

Schoen, D.J. and A.H.D. Brown1995. Maximising genetic diversity in core collections of wild relatives of crop species, p. 55–76. In: T. Hodgkin, A.H.D. Brown, T.J.L. van Hintum, and E.A.V. Morales (eds.). Core collections of plant genetic resources. Intl. Plant Genet. Resources Inst./Wiley-Sayce, Rome.

Tanksley, S.D. and S.R. McCouch. 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. Science 277:1063–1066.

Upadhyaya, H.D. and R. Ortiz. 2001. A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theor. Appl. Genet. 102:1292–1298.

U.S. Department of Agriculture. 2004. National genetic resources program. Germplasm resources information network (GRIN). 23 Sept. 2008. <http://www.ars-grin.gov/cgi-bin/npgs/html/desclist.pl?115>.

van Hintum, T.J.L. 1999. The core selector, a system to generate representative selections of germplasm collections. Plant Genet. Resour. Newsl. 118:64–67.

Volk, G.M., C.M. Richards, A.A. Reilley, A.D. Henk, P.L. Forsline, and H.S. Aldwinckle. 2005. Ex situ conservation of vegetatively propagated species: Development of a see d-based core collection for *Malus sieversii.* J. Amer. Soc. Hort. Sci. 130:203–210.

Wang, J.C., J. Hu, H.M. Xu, and S. Zhang. 2007. A strategy on constructing core collections by least distance stepwise sampling. Theor. Appl. Genet. 115:1–8.