EXPLORATION OF GENOMIC-BASED STRATEGIES FOR SCREENING AND SELECTION FROM A CARROT *(DAUCUS CAROTA)* GERMPLASM COLLECTION

Keo Corak

A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science
in Agroecology

at the
UNIVERSITY OF WISCONSIN-MADISON
2018

Table of Contents

Chapter 1: The use of diverse germplasm collections to meet breeding goals important to organic agriculture using carrot (*Daucus carota*) as a model crop

Chapter 2: Association analysis reveals the importance of the *Or* gne in carrot (*Daucus carota* L.) carotenoid accumulation and domestication

Chapter 3: Comparison of representative and "custom" methods of generating core subsets of a carrot (*Daucus carota)* germplasm collection

List of Tables

Chapter 3

List of Figures

Chapter 2

Chapter 3

List of Appendices

Chapter One:

The use of diverse germplasm collections to meet breeding goals important to organic agriculture using carrot (*Daucus carota*) as a model crop

Abstract:

This work takes up an old question of renewed importance to breeding programs. When faced with a large, genetically diverse group of crop accessions for which there is only imperfect and incomplete data, how does one go about making strategic choices regarding which accessions to prioritize? Separate analyses addressing this question compose chapters two and three of this thesis. In this chapter, I argue that developing vegetable varieties for organic agriculture in the United States presents a timely test situation in which to explore this issue. Traits important for success in organic environments and markets have not been prioritized in conventional breeding programs. Researchers need to turn toward genetically diverse collections to incorporate desired traits into breeding populations. These genetic resource collections hold great potential for breeding programs, but are underutilized. Carrot, a horticulturally important species with moderate genetic resources, allows us to explore strategies to choose accessions in germplasm collections in a manner extensible to other crops.

Plant Breeding for Organic Systems: Overview and Rational

There is a growing interest in sustainable agricultural systems; those that seek to promote ecological balance, to conserve natural resources and biodiversity, and to reduce the use of off-farm inputs through the use of varied mechanical and biological strategies (USDA-SARE). The strategies used are as diverse as the farms that practice them, but key similarities can be made. Often, synthetic chemical herbicides and insecticides are not used. Instead different management strategies such as mechanical control or intercropping are used to control pests. Rotational cropping systems, including cover cropping or green manures, may be used to maintain and restore fertility. Many of these strategies are multifunctional; rotational cropping, for example, can both add nutrients to the soil and reduce pest pressures over time (Lin 2011). In the United States, national guidelines for organic production provide rules and regulations regarding these systems (NOP-ARS).

Both farmers and consumers are drawn to organic and sustainable agriculture. Since 2000, organic acreage has more than tripled in the United States (USDA ERS 2013). At the same time, revenues from organic food sales have increased by 500% (Lernoud and Willer 2017). Farmers receive a price premium for certified organic produce; this reflects not only the higher labor costs typically associated with organic production but also consumer demand for organic products (OTA 2015).

Due to the increased interest in organic and sustainable farming practices, farmers, researchers and extension agents have been focused on developing management strategies suitable for organic farming conditions. In contrast, there has been relatively little attention paid to crop breeding specifically for organic environments. This is concerning because crop varieties that were bred for conventional high-input farming systems do not necessarily perform well under organic management.

Because organic and low-input farmers have different production challenges than conventional farmers, they require unique breeding solutions. Lammerts van Buren *et al*

(2009) found that not only do many conventionally-bred varieties lack traits necessary for success in organic systems, certain traits such as semi-dwarfism in conventional varieties are actually counter-productive in organic systems. Strong genotype-by-environment (GxE) interactions characterize differences in performance between organic and conventional environments. In a study of wheat genotypes the varieties that performed the best in conventional systems were not the same as the best-performing lines in organic systems (Murphy et al 2007). These results suggest organic agriculture, rather than performing uniformly lower than conventional agriculture, is instead highly dependent on choice of appropriate cultivar. Organic farmers therefore need crop varieties that are resistant to diseases and pests, efficient at using nutrients and are specifically bred for their production environments (Woolfe et al 2008, Hultengren et al 2016).

Organic farmers also require crop varieties suited to organic markets, which can differ substantially from conventional markets. Organic growers often contract with gourmet restaurants and specialty markets in which the visual appeal and flavor of their produce is paramount. At farmers markets as well, novel crop varieties catch the eye of shoppers. Crop varieties with novel shapes and colors, as well as high visual appeal and good flavor, are important for many organic growers.

Crop varieties bred for organic and low-input systems are insufficient, both in number and in kind, to meet current needs. By some estimates, 95% of the crop varieties grown under organic management were bred for conventional systems (Lammerts van Buren *et al* 2009). When surveyed, 72% of organic growers (n=54) in the Pacific Northwest agreed that there were crops in need of organic plant breeding.

87% of the same group agreed that varieties bred for organic production were important to the success of organic agriculture (Hubbard and Zystro 2016**).** Across all crops, disease resistance, yield and weed competiveness were highlighted as important traits. In the Northeast United States, a mix of organic and small-scale survey participants (n=344) indicated a need for organic vegetable varieties with improved storability, cold hardiness, disease resistance and flavor/appearance (Hultengren et a*l* 2016). A survey of 100 participants engaged in some aspect of organic plant production in Germany indicated that grain legumes, cabbages and oilseeds were in need of new varieties for organic production. This group targeted disease resistance, yield stability, open-pollination and flavor as key traits (Wilbois and Messmer 2015). Interestingly, in these studies, yield was not necessarily the most important trait for organic growers. While sufficient yield is important, and yield losses due to pest pressure are undesirable, many growers apparently prioritize other traits in deciding what to plant.

In order to meet the needs of organic growers for new varieties, differences in priority traits and crops across regions—which indicate both differences in markets and environments — will necessitate coordinated, decentralized breeding efforts (Desclaux 2005). These breeding efforts should take place in organic research plots and farms, although in some cases information from non-organic trials may be beneficial (Kamran et al 2014, Kokare et al 2014, Przystalksi et al 2008). In light of the high variability and presence of GxE interactions that characterize organic systems, Crespo-Herrera and Ortiz (2015) suggest treating different organic systems as their own target environment and selecting within them. Participatory and evolutionary plant breeding methods are being actively developed to meet the variety needs of diverse organic and low-input growers (Campenelli 2015, eOrganic 2014 (NOVIC), Phillips and Wolfe 2005, Chable et al 2008).

In order to continue to develop new varieties for organic growers, not only is it necessary to breed in organic environments, it is also important to increase the genetic diversity accessible to organic breeding programs (Lammerts van Buren et al 2005). Older landraces and unimproved cultivars may have traits advantageous is organic systems that have not been necessary to maintain in elite cultivars. These germplasm accessions are attractive candidates for parents in organic plant breeding programs. Additionally, high levels of environmental variability in organic systems can be buffered by high genetic diversity and phenotypic plasticity in crop varieties (COBRA). The efficient maintenance and use of genetic resources is therefore of paramount importance to organic agriculture.

Organic farm management differs from conventional management in significant ways. As acreage is converted to organic production across the United States and worldwide, new challenges emerge. The results of several farmer surveys indicated that organic vegetable varieties are insufficient to meet current needs. As climate systems continue to change and farmers are required to adapt their practices, new varieties appropriate to altered environments will need to be rapidly developed. It is therefore of vital importance to increase the use of genetic resources to increase diversity in breeding programs for organic systems.

Genetic resource collections use in organic breeding: promise and challenges

Genetic resource collections (GRC) are assemblages of living plant material – a kind of library of seeds, tissue cultures, or tubers – that function as "repositories of genetic variation" (Tanksley and McCouch 1997). These repositories are important sources of disease resistance, adaption to different environments and novel traits. Most GRC are maintained and used at the regional or country level. There are approximately 1750 such GRC housing over 7 million accessions worldwide (McCouch et al 2012). In addition, several large GRC have been managed by the Food and Agriculture Organization of the United Nations since 1994 and are regulated under International Treaty on Plant Genetic Resources for Food and Agriculture (FAO). GRC have been established for all major commodity crops as well as for many other species of horticultural and agronomic importance including: sorghum (Upadhyay et al 2009), soybean (Oliveria et al 2010), peanut (Dwivedi et al 2008), bread wheat (Balfourier et al 2007), cowpea (Mahalakshmi et al 2007) and many others. GRC typically include landraces, heirloom varieties and wild or weedy accessions of crop species.

Domestication and breeding often (but not always) result in a reduction of genetic diversity, therefore, the preservation of genetically diverse accessions in GRC is necessary to mitigate the undesirable effects of genetic erosion in crop species. Over time, the process of breeding with narrow sets of germplasm will tend to erode genetic variation as allelic variants become fixed in narrow populations through the action of selection or drift (Rief et al 2005, Lu and Bernardo 2001). Modern farming practices have shifted to favor uniform and high-yielding crop varieties, providing technical and economic incentives to minimize field crop genetic diversity (Plucknett and Smith 1987).

Crosses between genetically related, high-yielding parents can reliably produce offspring with improved yield traits but such crosses further narrow the genetic base of a breeding populations (Tanksley and McCouch 1997). In order to make continued crop improvement via selection, there must be genetic variability in genetic regions related to the trait of interest. When this is lacking in elite breeding pools, breeders can turn toward GRC as a source of new genetic variability. As such, GRC are especially appealing as sources of variation for traits of importance to organic and low-input agriculture because variation for these traits may be lacking in elite lines (Lammerts van Buren *et al* 2002, Ostergarde *et al* 2009).

A recent review by Byrne and colleagues outlines the recent uses of the germplasm accessions housed in the USDA-ARS National Plant Germplasm System (NPGS) by plant breeders (2017). The NPGS is a multi-location, multi-crop system of GRC that it relatively well funded compared to many other germplasm collections worldwide. It therefore should not be taken as representative of the ways in which *all* GRC are currently being used, but because of its size and resources, it provides an interesting model to explore potential use and/or improvement of existing GRC. Many of the uses cited in this review related to disease and insect resistance, as noted by the authors. Discovery of wild and landrace sources of resistance to late blight (tomato), Russian wheat aphid (wheat), downy mildew (spinach) and their incorporation into breeding programs are notable recent successful uses of NPGS resources. The use of GRS to improve quality traits and yield has been less significant. However, Tanskley and McCouch (1997) argue that advantageous allelic variants that increase yield are present in wild accessions, but are often hidden in genetic backgrounds that limit their discovery. Their use of backcrossed introgression lines (BILs) identified several wild QTL that increased rice yield when introgressed into elite backgrounds.

In addition to genetic diversity for specific traits, GRC accessions also have advantageous patterns of genome-wide diversity. Reynolds et al (2007) confirmed that landraces housed in GRC can be both genetically distant from elite lines as well as other landraces. These landraces have high genetic diversity both within and between lines (Mayer et al 2017, Mazzucato et al 2007). Shorter haplotype blocks associated with increased recombination events compared to mapping populations can facilitate association studies if population structure is known or estimated. These features of the genetic diversity of GRC lend them to use both in functional genetic studies as well as breeding efforts (Bandillo et al 2017, Huang et al 2012, Gebhardt et al 2004).

Despite the creative and successful uses of GRC to identify both valuable accessions and genetic regions associated with traits of interest, there are still significant challenges associated with GRC that have precluded their widespread use. Some of the issues associated with maintaining and using GRC are quite technical. For example, the regeneration of germplasm within the collection to preserve seed viability will tend to reduce the genetic diversity within accessions, especially within cross-pollinating species (Cross and Wallace 1993, Parzies et al 2000). This can be mitigated somewhat by careful regeneration schemes, however it’s likely that preservation will tend to erode genetic diversity in the collection over time. Relatedly, *ex situ* GRC have been likened to storehouses or museums for germplasm material and critiqued for failing to capture dynamic genetic diversity (Peres 2016). These issues deserve attention, but are largely outside the purview of this chapter.

On the other hand, a major reason for the underutilization of GRC in plant breeding efforts is that breeders have difficulty determining which accessions will be useful to incorporate into their programs. Because GRC can be very large in size (several thousand accessions for many cereal grains) exhaustive evaluation of the materials in them can be nearly impossible (McCouch et al 2012). Field phenotyping for traits like plant height or yield is expensive, time consuming, and not necessarily predictive of performance in environments other than the one tested due to GxE interactions (Plucknett and Smith 1987). Measured phenotypes may not accurately represent the potential utility of a given accession, especially if valuable alleles are “hidden” in unadapted accessions. Therefore, data on each accession is often limited and may only include information about geographic origin and basic morphological traits such as seed coat or market class: traits which do not necessarily help researcher predict the performance of accessions (Jansky, Dawson and Spooner 2015).

Genomic resource collections are important sources of functional genetic diversity and can be used to facilitate gene discovery, however their use is under-realized. To improve the use of GRC in plant breeding, methods to better identify relevant accessions from GRC are required. With this goal in mind, we seek to respond to the challenges associated with incorporating diverse accessions from GRC into plant breeding programs using carrot as a model crop.

Carrot as a model GRC

The two studies included in this thesis use a medium-sized collection of cultivated and wild carrot germplasm as a model in which to explore methods of identified interesting and relevant accessions from GRC. The use of carrot as a model crop is justified because of its economic and nutritional importance, physiology/reproductive biology, breeding history, genetic structure and the genetic resources that are available to leverage for breeding and gene discovery. Carrot is an important vegetable crop with increasing genetic resources, however it has received relatively little breeding attention. Therefore, not only is it an appropriate choice of species for this study, insights gleaned will also aid in its further improvement.

Carrot is economically and nutritionally important crop. Grown both for fresh and processing markets, it has a US farm gate value of 820.4M USD (USDA-NASS, 2016). The largest source of provitamin A in the US diet, carrot is a highly nutritious vegetable and is also highly palatable to consumers (Simon et al 2009). Modern breeding has dramatically increased both sugar and carotenoid content of elite carrot varieties. However, many of these cultivars are susceptible to pest and disease. They also have slow seed germination and poor early top growth. While carrot is an important crop for organic and small market growers, less research attention has been paid to improving these traits -- which are important to the success of carrot cultivation in small-scale and low-input systems.

Modern cultivated carrot is a biennial diploid species (2n=18) (Stein 1994). Seed production follows a requisite 6-8 week vernalization of the carrot taproot, which allows it to be cultivated as an annual if seed production can take place in a winter nursery or greenhouse (Simon and Goldman 2007). It reproduces primarily through outcrossing with a high reproductive capacity. Accordingly, it suffers severe inbreeding depression.

Carrot was domesticated in Central Asia around 5000 years ago and was subsequently brought both east and west into Europe and East Asia (Banga 1957, Baranski 2012). While wild carrot is endemic to all three of these regions, genetic and historical analysis points to only a single domestication event (Iorizzo et al 2013). Commercial breeding in the West has focused on orange color since the 1600s, but greater diversity in color and root morphology exist in Eastern types (Stolarcyk and Janick 2011). Unlike many domesticated species, which have markedly reduced genetic diversity compared to their wild progenitors, Iorizzo et al (2013) showed that there is little reduction of genetic diversity in cultivated carrot compared to wild accessions. This is likely because carrot freely outcrosses with its wild relatives. Furthermore, a genetic study of commercially cultivated carrot varieties in the US suggested that these varieties form one large breeding pool with moderate genetic diversity and found that there is no significant subgroup differentiation along color or market class (Luby et al 2016).

We draw on diverse germplasm housed in several different carrot collections to inform the descriptive and analytical work that comprises chapters two and three. The cultivated accessions within the carrot USDA National Plant Germplasm (USDA-NPGS) have been phenotyped and genotyped by genotype-by-sequencing (GBS) as part of a related project. Additionally, we include data from a well-characterized collection of 170 open-pollinated carrot varieties (Theisen et al 2016) and from the Luby et al (2016) study of commercial carrot cultivars.

Within carrot GRC, there is a is significant morphological and genetic diversity for key agronomic and quality traits, however this diversity is underutilized both in research and in breeding programs. While research using mapping population to interrogate genetic regions underlying important market traits is ongoing, progress – which is reviewed in the following section – has been slow. The recent publication of a partially annotated carrot genome (Iorizzo 2016) and the increasing affordability of high-density molecular markers, however, should increase the pace of gene discovery by allowing for association analysis and functional studies in diverse populations.

QTL analysis in carrot

There are two complementary approaches used to identify regions of the genome associated with a trait of interest: linkage analysis using experimental mapping populations and association mapping using diversity panels or natural populations. Both have been used to explore the genetic architecture of root color and other traits in carrot.

Quantitative trait loci (QTL) are sections of the genome associated with a specific trait. Analysis of quantitative trait loci through either linkage or association analysis allows researchers to elucidate the genetic regions associated with complex traits such as yield, horizontal disease resistance, height and others. Such traits are considered quantitative because they are controlled by a few genes with large effects and many genes with small effects. The segregation of different combinations of alleles in a population leads to an approximately normally distributed range of phenotypes.

In linkage analysis, populations are developed by crossing inbred parents with distinct phenotypes; F1 and later generation populations are expected to display a range of continuous variation in the trait of interest (Lynch and Walsh 1998, p 431). Polymorphic DNA markers, such as SNPs or SSRs, that differentiate the parents and segregate in the F1 population are identified and mapped. Markers that lie close to a given QTL will not segregate independently of the QTL and can be used to statistically associate a given phenotype with a specific genetic region. Then, the contribution of each QTL to the total variation in the trait can be estimated using additional techniques (Bernardo 2014, p185).

Precision of linkage analysis is limited by the frequency of recombination in experimental populations and is further dependent on the heritability of the trait, QTL effect size, and the presence of multiple QTL on the same chromosome (Bernardo 2014, 184). Also, when analysis is performed in populations that have undergone only a few generations of recombination, large haplotype blocks will persist in the genotyped progeny. Markers on a given haplotype block will be statistically associated with a given trait even if they lie far from the causative genetic region. QTL discovered in one mapping population are not always found in others, suggesting both a) complexity of genetic conditioning of phenotypes and b) background genetic effects that influence expression of phenotypes. Despite these limitations, in combination with other approaches linkage analysis has been successfully used to detect and identify genes that condition complex traits.

A second approach to identifying genetic regions associated with complex traits is known as association analysis (AA). Association analysis takes advantage of historical recombination in natural populations and is made possible through high-density genetic maps that allow for estimation of the decay of linkage disequilibrium (LD) across the genome. Linkage disequilibrium refers to the non-random and reduced recombination of specific alleles i.e. alleles in high LD occur more frequently together than is expected by chance (Hartl and Clark 1997, p. 95). With sufficient historical recombination, LD decays rapidly across the genome; therefore, the power to associate a narrow genetic region with a trait of interest is typically higher than in linkage analysis. Physical proximity, however, is not the only phenomenon that results in LD. Distinct population structure and/or relatedness between lines, for example, can lead to LD between unlinked loci resulting in spurious associations. Such structure can be suitably accounted for in association models (Bernardo 2014, p. 187). For these reasons, AA is emerging as a powerful alternative to traditional linkage mapping.

In carrot, linkage analysis has primarily been used to explore the genetic architecture of root color. Carotenoids accumulate in the carrot root and are responsible for their diverse yellow and orange coloration. Because carotenoids play an important role in human nutrition, understanding the genetic control of their synthesis and accumulation in the carrot root has proven to be an important breeding goal. As early as 1979, Bushland and Gableman proposed a two-locus genetic model for the accumulation of color in carrot. They found that a dominant locus Y conditioned a white coloration and that the homozygous recessive genotype produced an orange coloration. A second locus, Y2, conditioned yellow color with the homozygous recessive again developing an orange root phenotype. The model was confirmed by QTL analysis of two unrelated populations segregating for root color. In separate crosses of orange to white and orange to dark orange inbred carrot lines, Santos and Simon (2002) detected two major clusters of QTL conditioning root color, consistent with the model proposed by Bushland and Gableman. These clusters were later mapped onto chromosomes 5 (Y) and 7 (Y2) (Just et al 2009, Cavagnaro et al 2011). Ellison et al (2017) localized a single major QTL for beta-carotene on chromosome 7 which overlapped Y2. Study of the recently published carrot genome has suggested a candidate for the Y locus on chromosome 5, DCAR_032551 which may regulate carotenoid accumulation in roots by conditioning expression of the necessary precursors to the carotenoid biosynthetic pathway (Iorizzio et al, 2016). An alternative hypothesis stemming from an association analysis study in a broad unstructured discovery panel of carrots posits carotenoid biosynthetic genes YEP and PDS as candidates for the Y2 and Y loci, respectively (Jourdan et al 2015).

Purple carrots actually precede orange carrots in the domestication record and are still common in parts of the world but have received considerably less attention from breeders. Due to their unique color and nutrition profile, however, interest in the purple carrots has been renewed. Yildiz et al (2013) mapped known anthocyanin biosynthesis genes in a population that segregates for the P1 locus that conditions purple color. They found that P1 mapped to chromosome 3 and that two of eight known anthocyanin genes were linked to P1. They also found that increased transcription of these genes was positively associated with anthocyanin accumulation. Cavagnaro et al (2014) developed the first SNP based linkage map in carrot and used it to study the accumulation of anthocyanins in roots and petioles. They suggested two and one gene models for purple color in the root and petiole, respectively. Eight QTL conditioning purple color in two clusters on chromosome three were identified. Many of these QTL co-localized with QTL for anthocyanins.

Other traits in carrot have been studied to a lesser extent than root color. Ali et al. (2014) identified a source of resistance to root-knot nemotode M. javanica on chromosome eight in a mapping population of a cross between susceptible and resistant cultivars, Mj-2. Iorizzio et al. (2016) identified a resistance gene that co-localized with nematode resistance QTL Mj-1at a different locus on chromosome eight. Clerc et al. (2015) developed two connected populations from crosses of cultivars susceptible and resistant to leaf blight. They explored the stability of QTL detection across years and found evidence of 11 QTL conditioning blight resistance, however some of these were only detected in a single year, indicating the presence of QTLxE interactions. Limited exploration of genes conditioning fertility and vernalization has been undertaken by both Alassandro et al. (2012) and Bhudan et al. (2014). Alassandro found dominant single gene loci for early flowering time and restoration of CMS on chromosomes 2 and 9. Bhudan et al. performed a cross of flowering mutants but found no significant associations suggesting further work on the genetic architecture of flowering and germination is needed.

In carrot, linkage analysis has been primarily applied to explore genetic control of root color. Work to genetically characterize carrot GRC should facilitate association analyses of under-studied traits that could be useful in breeding.

Population structure, selection signatures and association analysis in carrot

To improve the use of GRC for breeding in carrot, it is necessary to understand the genetic diversity of the species, which in crop species is shaped by domestication and continued selection. In the second chapter of this thesis, we perform complementary analyses to interrogate relationships between domestication and changes in diversity across the carrot genome. Using a large dataset of domesticated and wild carrot, we survey the carrot genome for signatures of selection and look for genetic associations with domestication phenotypes. While auxiliary to the main theme of this thesis – that is, the optimization of genetic resource use for specific breeding goals – the results presented in the second chapter complement those aims nicely. Because we use a diverse set of germplasm in our analysis, we move our understanding of carrot domestication forward in a way that has not been achieved using smaller datasets. Not only are diverse GRC important for breeders, they are also interesting populations to study in and of themselves.

Previous reports of population structure in carrot suggest a genetic separation between Eastern and Western accessions but otherwise little reduction in diversity when comparing cultivated to wild carrot (Iorizzo, 2013). Domestication is generally accompanied by a significant genetic bottleneck but this does not seem to have occurred in the domestication history of carrot, likely because it freely outcrosses with wild *Daucus* worldwide.

We use several different methods; genome-wide LD, STRUCTURE, principal component analysis (PCA), phylogeny, pairwise F_{st}, and expected heterozygosity on a wide set of carrot germplasm to deepen our understanding of carrot population structure.

Linkage disequilibrium is influenced in predictable ways by many genetic processes. Comparing LD and the rate of LD decay across the genome between subgroups can signal the extent to which selection, recombination rate, genetic drift, mating system, population structure and linkage structure the genome. STRUCTURE (Prichard et al, 2000) is a standard procedure used to model the number and composition of subpopulations within a larger population. It uses an iterative process to assign individuals to subpopulations that individually meet the expectation of HW equilibrium. In chapter 2 we complement the results from STRUCTURE with principal component analysis (PCA) to better visualize the genetic distances between apparent STRUCTURE groups via the principal component scores of individuals. PCA is an appropriate way to characterize genetic structure in populations and relies on fewer assumptions than STRUCTURE (Odong, 2011).

We also compute pairwise F_{st} and expected heterozygosity between and within each STRUCTURE group, respectively. In structured populations, heterozygosity is lower than expected under HW equilibrium due to inbreeding within subpopulations; F_{st} quantifies the extent to which expected heterozygosity is reduced compared to predicted in order to suggest the extent of population differentiation into subgroups. (Wright 1951, Nei 1973, Weir and Cockerham 1984).

We conduct a genome wide association analysis to detect regions of the carrot genome associated with a phenotype thought to be associated with selection after domestication; orange color. Orange is the most common color in western carrot and was heavily selected for in the 16th century (Simon, 2000). We hypothesize that our large dataset of improved cultivars, historic cultivars and wild accessions would allow us to identify regions of the genome putatively controlling orange carrot root color that may not have been identified in previous QTL studies. Because our GWAS is performed in a diverse population, it can detect genomic variants associated with traits important for domestication or selection that may be fixed in domesticated populations. We leverage results from the population structure analysis to control spurious associations in our GWAS results, which can result from misclassifying genetic differentiation due to population substructure. Resequenced lines are used to identify a putatively causative SNP in a region significantly associated with carrot color.

To explore the hypothesized link between orange root color and artificial selection in the population history of carrot we look for signatures of selection in the genome, comparing orange and non-orange domesticated carrots. When selection acts to increase the frequency of a specific allelic variant in the genome there is often a concomitant reduction in genetic diversity around the variant within a population because nearby regions will tend to be inherited with it (Akey et al, 2002). This will tend to increase genetic differentiation among subpopulations carrying different versions of the allele. These genomic phenomena, known as selective sweeps, can be detected by calculating F_{st}, nucleotide diversity and other metrics of population differentiation across the genome. Selection may be acting in regions with enhanced levels of differentiation. Comparing results of our GWAS and selective sweep analyses allow us to identify a candidate gene target for selection on orange root color.

Understanding the genetic structure of a population is necessary for many reasons. In chapter two, describing the genetic structure of a diverse carrot collection informs our analysis and interpretation of GWAS and selective sweep results. Within larger breeding contexts, it allows us to define breeding pools and to identify priority areas for conservation.

Core collections

In the third chapter of this thesis, we revisit the question of effective use of GRC materials. Using the carrot collection, we explore newly emerging strategies that leverage reduced-representations sequencing data to sample relevant accessions. The following section describes the theoretical concepts and practical considerations related to sampling from GRC that are relevant to our current work with carrot.

It has long been recognized that the large size of GRC is both a benefit and a limitation. Frankel (1984) first observed that as germplasm collections grew in size over time, it would become progressively more challenging to maintain and catalog all accessions within them. New management strategies would therefore be required: ones that allowed for fewer number of accessions to be prioritized for evaluations. From these observations, the concept of a “core collection” emerged.

Drawing on principals from theoretical genetics, Brown showed that for a general case, it was possible to construct a “core collection” which would maintain the allelic diversity of a collection in a markedly reduced number of samples. This core collection would represent “with a minimum of repetiveness, the genetic diversity of a crop species” (Frankel 1984). Using an infinite neutral alleles model, Brown showed that 70% of rare alleles in a single population would be preserved in a sample of just 10% of the total collection (Brown 1989b). If desirable alleles are uniformly dispersed in a population, a core collection could then be formed simply by randomly sampling individuals. In many cases, however, allelic diversity is non-uniformly distributed across a collection i.e. there are rare alleles localized to a specific subgroup. To increase the chances of including these types of alleles in a core, the collection should first be stratified into subgroups of even within-group allelic diversity, from which samples could be chosen.

To develop a core collection via stratified sampling, three different aspects of sampling must be considered 1) which variables should be used to classify individuals into groups 2) how accessions should be sampled from within those stratified groups and 3) how the representativeness of the core to the whole collections will be measured (Odong 2013).

1) Stratification

Many different variables could theoretically be used to stratify a collection. Common ones include geographic origin, morphological descriptors, quantitative phenotypic traits and, increasingly, molecular genetic data. The choice of a particular variable or set of variables has often depended on the data that already exists on accessions within a collection. Each strategy has both benefits and limitations.

Grouping individuals by geographic region of origin makes intuitive sense and can be easily accomplished for most collections. As such, it is one of the most widely used in the literature on core collections (Mahalakshmi et al 2007, Jewell et al 2012, Malosetti et al 2001, Dwiveldi et al 2008, Upadhyaya et al 2003, Igartua et al 1998). Based on the assumption that geographically separate individuals will be more distantly related than those in close proximity to one another, this strategy can prevent oversampling of closely related individuals from regions with many representative accessions in the collection. However, the assumption of isolation-by-distance may not be a valid one in all cases (Ghislain 2006, Skroch 1998). Alternatively, grouping by political borders may miss finer ecological gradations distinguishing related subgroups of individuals (Brown 1989).

Morphological traits can be further used to subdivide groups. While data on entries within a collection is often incomplete, easily observable and highly heritable traits like seed type and root color are often recorded and used to construct cores (Masoletti et al 2001, Upadhyaya et al 2003, Huaman et al 1999, Zewdie et al 2004). For traits controlled by a few major loci, however, morphological features may be a poor measure of overall allelic diversity**.**

While geographic origin and morphological descriptors facilitate the grouping of accessions into discrete subgroups, it is less straightforward to use quantitative phenotypic traits to group individuals. Generally, some estimate of distance between individuals must be made using standardized datasets and then either agglomerative clustering techniques or dimensionality reduction analysis are used to group accessions. There are several examples of cores generated using phenotypic and evaluation data (Diwan et al 1995, Rodino et al 2003, Tai and Miller 2001) however phenotypic data is generally not the only descriptor used to subdivide a collection.

Odong (2011) has shown that hierarchical clustering methods are appropriate for molecular marker data when there are genetically distinct subgroups. Balfourier et al (2007), Chavarriaga-Aguirre et al (1999), Erskine et al (1991), Xiurong et al (2000), Belaj et al (2012), Xu et al (2016) and others have used molecular marker data to group accessions. While guidelines exist, the question of an appropriate number of subclusters is not straightforward analytically. Sometimes subjective interpretation distinguishes obvious genetic subgroups, but this is not always the case. The choice of maker also likely influences cluster analysis- neutral markers may adequately describe the overall genetic diversity of a collection however they may not capture functional diversity for key agronomic traits.

2) Sampling

Once accessions within a collection have been stratified into the desired subclusters, samples from each subcluster are combined to form the core. Because subclusters are often quite different in size, various methods to determine the number of individuals to sample per cluster have been proposed. Choosing a constant number from each cluster will insure individuals from small clusters are included in the core but may under sample large clusters. Choosing a sample proportionate to the size of a cluster may under-sample small clusters while choosing a sample based on the logarithm of the cluster size balances the other two approaches (Brown 1989). Other methods have been proposed as well (see Franco et al 2005, Hu et al 2000). A second class of methods based on the optimization of different evaluation measures does not initially stratify a collection at all (Thachuk 2009).

3) Evaluation

A core collection should represent the diversity of the entire collection. Evaluating the diversity of the core should ideally be considered not only along the variables used to stratify the collection but by other metrics as well. For example, if geographic origin data is used to stratify a collection into subgroups, both the geographic and morphologic diversity of the core should be evaluated if possible. Odong (2013) treats the evaluation of core collections in detail. For morphologic, phenotypic and geographic data, the following methods can be used to compare diversity between the core and the entire collection: summary statistics, principal component analysis, diversity indices, class coverage and goodness of fit tests. Genetic diversity measures can be used in cases where robust genotypic information is available.

In 1989, Brown wrote about using accessions in a core collection as a “guide” to other, better individuals held in the whole collection. Lacking from the evaluation criteria applied to most core collections, however, is an analysis of the predictive ability of the core i.e an analysis of the extent to which information about material in a core collection can be useful to identify similarly useful accessions outside of the core. To update Brown’s language, we can state that a good core collection should have sufficient power to predict traits in other accessions. It is unlikely, however that the variables commonly used to construct cores will be helpful in this aim (Spooner et al 2017). Emerging research seeks to evaluate if predictive methods based on the genetic relationships between individuals are now robust to be used to construct cores with higher predictive ability (Crossa et al 2016, Gorjanc et al 2016).

In theory, developing a core collection allows researchers to reduce the working size of a collection without losing significant information about the diversity in the entire collection. However, in practice such a goal is fraught with technical challenges. New methods to developing useful subsets of germplasm with high predictive ability are required if GRC are to be used to their full potential.

Conclusion:

To better serve organic and low-input farmers, breeders need to develop crop varieties suited to unique farming conditions. Because crop traits important to success in organic environments have been de-prioritized in many conventional programs, GRC are attractive sources of novel germplasm that can be introgressed into elite lines. The large size of GRC, however, makes it challenging to collect and curate comprehensive data on accessions, which retards breeding efforts. In this study, we explore the genetic diversity in a collection of carrot accessions and use this collection to compare strategies of choosing representative subsets of accessions from a collection. We expect that our findings in carrot, an important vegetable crop with moderate genetic resources, will be of interest to breeders of other crops interested in introgressing diverse material into breeding programs.

Literature Cited

Akey, J. M., Zhang, G., Zhang, K., Jin, L., & Shriver, M. D. (2002). Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, *12*(12), 1805–1814. https://doi.org/10.1101/gr.631202

Alessandro, M. S., Galmarini, C. R., Iorizzo, M., & Simon, P. W. (2012). Molecular mapping of vernalization requirement and fertility restoration genes in carrot. *Theoretical and Applied Genetics*, *126*(2), 415–423. https://doi.org/10.1007/s00122-012-1989-1

Arber, W. (1984). *Genetic Manipulation: Impact on Man and Society*. Cambridge University Press.

Background | Cobra Div. (n.d.). Retrieved January 11, 2018, from http://www.cobra-div.eu/background/

Balfourier, F., Roussel, V., Strelchenko, P., Exbrayat-Vinson, F., Sourdille, P., Boutet, G., … Charmet, G. (2007). A worldwide bread wheat core collection arrayed in a 384-well plate. *Theoretical and Applied Genetics*, *114*(7), 1265–1275. https://doi.org/10.1007/s00122-007-0517-1

Banga, O. (1957). Origin of the European cultivated carrot. *Euphytica*, *6*(1), 54–63. https://doi.org/10.1007/BF00179518

Baranski, R., Maksylewicz-Kaul, A., Nothnagel, T., Cavagnaro, P., Simon, P., & Grzebelus, D. (2012). Genetic diversity of carrot (Daucus carota L.) cultivars revealed by analysis of SSR loci. *Genet Resour Crop Ev.*, *59*, 163–170.

Belaj, A., Dominguez-García, M. del C., Atienza, S. G., Urdíroz, N. M., Rosa, R. D. la, Satovic, Z., … Río, C. D. (2012). Developing a core collection of olive (<Emphasis Type="Italic">Olea europaea</Emphasis> L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genetics & Genomes*, *8*(2), 365–378. https://doi.org/10.1007/s11295-011-0447-6

Brown, A. H. D. (1989). Core collections: A practical approach to genetic resource management. *Genome*, *31*(2), 818–824.

Brown, A. H. D. (n.d.). The case for core collections. In *The Use of Plant Genetic Resources,* (pp. 136–156). Cambridge University Press.

Budahn, H., Barański, R., Grzebelus, D., Kiełkowska, A., Straka, P., Metge, K., … Nothnagel, T. (2014). Mapping genes governing flower architecture and pollen development in a double mutant population of carrot. *Plant Genetics and Genomics*, *5*, 504. https://doi.org/10.3389/fpls.2014.00504

Buishand, J. G., & Gabelman, W. H. (1979). Investigations on the inheritance of color and carotenoid content in phloem and xylem of carrot roots (Daucus carota L.). *Euphytica*, *28*(3), 611–632. https://doi.org/10.1007/BF00038928

Byrne, P. F., Volk, G. M., Gardner, C., Gore, M. A., Simon, P. W., & Smith, S. (2018). Sustaining the Future of Plant Breeding: The Critical Role of the USDA-ARS National Plant Germplasm System. *Crop Science*, *58*(2), 451–468. https://doi.org/10.2135/cropsci2017.05.0303

Campanelli, G., Acciarri, N., Campion, B., Delvecchio, S., Leteo, F., Fusari, F., … Ceccarelli, S. (2015). Participatory tomato breeding for organic conditions in Italy. *Euphytica*, *204*(1), 179–197. https://doi.org/10.1007/s10681-015-1362-y

Cavagnaro, P. F., Iorizzo, M., Yildiz, M., Senalik, D., Parsons, J., Ellison, S., & Simon, P. W. (2014). A gene-derived SNP-based high resolution linkage map of carrot including the location of QTL conditioning root and leaf anthocyanin pigmentation. *BMC Genomics*, *15*, 1118. https://doi.org/10.1186/1471-2164-15-1118

Chable, V., Conseil, M., Serpolay, E., & Lagadec, F. L. (2008). Organic varieties for cauliflowers and cabbages in Brittany: from genetic resources to participatory plant breeding. *Euphytica*, *164*(2), 521–529. https://doi.org/10.1007/s10681-008-9749-7

Chavarriaga-Aguirre, P., Maya, M. M., Tohme, J., Duque, M. C., Iglesias, C., Bonierbale, M. W., … Kochert, G. (1999). Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. *Molecular Breeding*, *5*(3), 263–273. https://doi.org/10.1023/A:1009627231450

Clerc, V. L., Marques, S., Suel, A., Huet, S., Hamama, L., Voisine, L., … Briard, M. (2015). QTL mapping of carrot resistance to leaf blight with connected populations: stability across years and consequences for breeding. *Theoretical and Applied Genetics*, *128*(11), 2177–2187. https://doi.org/10.1007/s00122-015-2576-z

Crespo-Herrera, L. A., & Ortiz, R. (2015). Plant breeding for organic agriculture: something new? *Agriculture & Food Security*, *4*, 25. https://doi.org/10.1186/s40066-015-0045-1

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., … Singh, S. (2016). Genomic Prediction of Gene Bank Wheat Landraces. *G3: Genes|Genomes|Genetics*, *6*(7), 1819–1834. https://doi.org/10.1534/g3.116.029637

Desclaux, D. (2018). Participatory Plant Breeding Methods for Organic Cereals : Review and Perspectives.

Diwan, N., McIntosh, M. S., & Bauchan, G. R. (1995). Methods of developing a core collection of annual <Emphasis Type="Italic">Medicago</Emphasis> species. *Theoretical and Applied Genetics*, *90*(6), 755–761. https://doi.org/10.1007/BF00222008

Dwivedi, S. L., Puppala, N., Upadhyaya, H. D., Manivannan, N., & Singh, S. (2008). Developing a Core Collection of Peanut Specific to Valencia Market Type. *Crop Science*, *48*(2), 625. https://doi.org/10.2135/cropsci2007.04.0240

Ellison, S., Senalik, D., Bostan, H., Iorizzo, M., & Simon, P. (2017). Fine Mapping, Transcriptome Analysis, and Marker Development for Y2, the Gene That Conditions β-Carotene Accumulation in Carrot (Daucus carota L.). *G3: Genes, Genomes, Genetics*, *7*, 2665–2675.

eOrganic. (n.d.). Retrieved January 11, 2018, from http://eorganic.info/novic/

Erskine, W., & Muehlbauer, F. J. (1991). Allozyme and morphological variability, outcrossing rate and core collection formation in lentil germplasm. *Theoretical and Applied Genetics*, *83*(1), 119–125. https://doi.org/10.1007/BF00229234

Franco, J., Crossa, J., Taba, S., & Shands, H. (2005). A Sampling Strategy for Conserving Genetic Diversity when Forming Core Subsets. *Crop Science*, *45*(3), 1035. https://doi.org/10.2135/cropsci2004.0292

Genetic Vulnerability of Major Crops. (1972). National Academies.

Ghislain, M., Andrade, D., Rodríguez, F., Hijmans, R. J., & Spooner, D. M. (2006). Genetic analysis of the cultivated potato Solanum tuberosum L. Phureja Group using RAPDs and nuclear SSRs. *Theoretical and Applied Genetics*, *113*(8), 1515–1527. https://doi.org/10.1007/s00122-006-0399-7

Gorjanc, G., Jenko, J., J. Hearne, S., & M. Hickey, J. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics*, *17*, 30. https://doi.org/10.1186/s12864-015-2345-z

Hu, J., Zhu, J., & Xu, H. M. (2000). Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theoretical and Applied Genetics*, *101*(1–2), 264–268.

Huamán, Z., Aguilar, C., & Ortiz, R. (1999). Selecting a Peruvian sweetpotato core collection on the basis of morphological, eco-geographical, and disease and pest reaction data. *Theoretical and Applied Genetics*, *98*(5), 840–844. https://doi.org/10.1007/s001220051142

Hubbard, K., & Zystro, J. (n.d.). State of Organic Seed. Organic Seed Alliance. Retrieved from http://seedalliance.org/ publications.

Hultengren, R. L., Glos, M., & Mazourek. (n.d.). *Breeding, Research, and Education Needs Assessment for Organic Vegetable Growers in the Northeast*. Retrieved from https://ecommons.cornell.edu/bitstream/handle/1813/44636/Breeding_Research_Education%20_Northeast.pdf?sequence=9&isAllowed=y

Igartua, E., Gracia, M. P., Lasa, J. M., Medina, B., Molina-Cano, J. L., Montoya, J. L., & Romagosa, I. (1998). The Spanish barley core collection. *Genetic Resources and Crop Evolution*, *45*(5), 475–481. https://doi.org/10.1023/A:1008662515059

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., … Simon, P. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics*, *48*(6), 657–666. https://doi.org/10.1038/ng.3565

Iorizzo, M., Senalik, D. A., Ellison, S. L., Grzebelus, D., Cavagnaro, P. F., Allender, C., … Simon, P. W. (2013). Genetic structure and domestication of carrot (Daucus carota subsp. sativus) (Apiaceae). *American Journal of Botany*, *100*(5), 930–938. https://doi.org/10.3732/ajb.1300055

Jansky, S. H., Dawson, J., & Spooner, D. M. (2015). How do we address the disconnect between genetic and morphological diversity in germplasm collections? *American Journal of Botany*, *102*(8), 1213–1215. https://doi.org/10.3732/ajb.1500203

Jewell, M. C., Zhou, Y., Loch, D. S., Godwin, I. D., & Lambrides, C. J. (2012). Maximizing Genetic, Morphological, and Geographic Diversity in a Core Collection of Australian Bermudagrass. *Crop Science*, *52*(2), 879. https://doi.org/10.2135/cropsci2011.09.0497

Jourdan, M., Gagné, S., Dubois-Laurent, C., Maghraoui, M., Huet, S., Suel, A., … Geoffriau, E. (2015). Carotenoid Content and Root Color of Cultivated Carrot: A Candidate-Gene Association Study Using an Original Broad Unstructured Population. *PLOS ONE*, *10*(1), e0116674. https://doi.org/10.1371/journal.pone.0116674

Just, B. J., Santos, C. A. F., Yandell, B. S., & Simon, P. W. (2009). Major QTL for carrot color are positionally associated with carotenoid biosynthetic genes and interact epistatically in a domesticated × wild carrot cross. *Theoretical and Applied Genetics*, *119*(7), 1155–1169. https://doi.org/10.1007/s00122-009-1117-z

Kamran, A., Kubota, H., Yang, R.-C., Randhawa, H. S., & Spaner, D. (2014). Relative performance of Canadian spring wheat cultivars under organic and conventional field conditions. *Euphytica*, *196*(1), 13–24. https://doi.org/10.1007/s10681-013-1010-3

Kokare, A., Legzdina, L., Beinarovica, I., Maliepaard, C., Niks, R. E., & Bueren, E. T. L. van. (2014). Performance of spring barley (Hordeum vulgare) varieties under organic and conventional conditions. *Euphytica*, *197*(2), 279–293. https://doi.org/10.1007/s10681-014-1066-8

Lammerts van Bueren, E. T., Jones, S. S., Tamm, L., Murphy, K. M., Myers, J. R., Leifert, C., & Messmer, M. M. (2011). The need to breed crop varieties suitable for organic farming, using wheat, tomato and broccoli as examples: A review. *NJAS - Wageningen Journal of Life Sciences*, *58*(3), 193–205. https://doi.org/10.1016/j.njas.2010.04.001

Lin, B. B. (2011). Resilience in Agriculture through Crop Diversification: Adaptive Management for Environmental Change. *BioScience*, *61*(3), 183–193. https://doi.org/10.1525/bio.2011.61.3.4

Lu, H., & Bernardo, R. (2001). Molecular marker diversity among current and historical maize inbreds. *Theoretical and Applied Genetics*, *103*(4), 613–617. https://doi.org/10.1007/PL00002917

Luby, C. H., Dawson, J. C., & Goldman, I. L. (2016). Assessment and Accessibility of Phenotypic and Genotypic Diversity of Carrot (Daucus carota L. var. sativus) Cultivars Commercially Available in the United States. *PLOS ONE*, *11*(12), e0167865. https://doi.org/10.1371/journal.pone.0167865

Mahalakshmi, V., Ng, Q., Lawson, M., & Ortiz, R. (2007). Cowpea [Vigna unguiculata (L.) Walp.] core collection defined by geographical, agronomical and botanical descriptors. *Plant Genetic Resources: Characterization and Utilization*, *5*(03), 113–119. https://doi.org/10.1017/S1479262107837166

Malosetti, M., & Abadie, T. (2001). Sampling strategy to develop a core collection of Uruguayan maize landraces based on morphological traits. *Genetic Resources and Crop Evolution*, *48*(4), 381–390.

Mayer, M., Unterseer, S., Bauer, E., de Leon, N., Ordas, B., & Schön, C.-C. (2017). Is there an optimum level of diversity in utilization of genetic resources? *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *130*(11), 2283–2295. https://doi.org/10.1007/s00122-017-2959-4

Mazzucato, A., Papa, R., Bitocchi, E., Mosconi, P., Nanni, L., Negri, V., … Veronesi, F. (2008). Genetic diversity, structure and marker-trait associations in a collection of Italian tomato (Solanum lycopersicum L.) landraces. *Theoretical and Applied Genetics*, *116*(5), 657–669. https://doi.org/10.1007/s00122-007-0699-6

McCouch, S. R., McNally, K. L., Wang, W., & Hamilton, R. S. (2012). Genomics of gene banks: A case study in rice. *American Journal of Botany*, *99*(2), 407–423. https://doi.org/10.3732/ajb.1100385

Murphy, K., Garland-Campbell, K., Lyon, S., & S. Jones, S. (2007). Evidence of varietal adaptation to organic farming systems. *Field Crops Research*, *102*, 172–177. https://doi.org/10.1016/j.fcr.2007.03.011

National Organic Program | Agricultural Marketing Service. (n.d.). Retrieved March 4, 2018, from https://www.ams.usda.gov/about-ams/programs-offices/national-organic-program

Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proc Natl Acad Sci USA*, *70*(12), 3321–3323.

Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. L. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theoretical and Applied Genetics*, *126*(2), 289–305. https://doi.org/10.1007/s00122-012-1971-y

Odong, T. L., van Heerwaarden, J., Jansen, J., van Hintum, T. J. L., & van Eeuwijk, F. A. (2011). Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? *Theoretical and Applied Genetics*, *123*(2), 195–205. https://doi.org/10.1007/s00122-011-1576-x

Oliveira, M. F., Nelson, R. L., Geraldi, I. O., Cruz, C. D., & de Toledo, J. F. F. (2010). Establishing a soybean germplasm core collection. *Field Crops Research*, *119*(2–3), 277–289. https://doi.org/10.1016/j.fcr.2010.07.021

Organic Trade Association (OTA). (n.d.). *Market Analysis*. Retrieved from https://www.ota.com

Phillips, S. L., & Woolfe, M. S. (2005). Evolutionary Plant Breeding. *Journal of Agricultural Science*, *143*, 245–254. https://doi.org/doi:10.1017/S0021859605005009

Plucknett, D., & Smith, N. (n.d.). *Gene Banks and the World's Food*.

Pritchard, J. K., Stevens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics Society of America*. Retrieved from http://web.stanford.edu/group/pritchardlab/publications/pdfs/PritchardEtAl00.pdf

Przystalski, M., Osman, A., Thiemt, E. M., Rolland, B., Ericson, L., Østergård, H., … Krajewski, P. (2008). Comparing the performance of cereal varieties in organic and non-organic cropping systems in different European countries. *Euphytica*, *163*(3), 417–433. https://doi.org/10.1007/s10681-008-9715-4

Reif, J. C., Zhang, P., Dreisigacker, S., Warburton, M. L., Ginkel, M. van, Hoisington, D., … Melchinger, A. E. (2005). Wheat genetic diversity trends during domestication and breeding. *Theoretical and Applied Genetics*, *110*(5), 859–864. https://doi.org/10.1007/s00122-004-1881-8

Reynolds, M., Dreccer, F., & Trethowan, R. (2007). Drought-adaptive traits derived from wheat wild relatives and landraces. *Journal of Experimental Botany*, *58*(2), 177–186. https://doi.org/10.1093/jxb/erl250

Rodiño, A. P., Santalla, M., Ron, A. M. D., & Singh, S. P. (2003). A core collection of common bean from the Iberian peninsula. *Euphytica*, *131*(2), 165–175. https://doi.org/10.1023/A:1023973309788

Santos, C., & Simon, P. (2002). QTL analyses reveal clustered loci for accumulation of major provitamin A carotenes and lycopene in carrot roots. *Molecular Genetics and Genomics*, *268*(1), 122–129. https://doi.org/10.1007/s00438-002-0735-9

Seufert, V., & Ramankutty, N. (2017). Many shades of gray—The context-dependent performance of organic agriculture. *Science Advances*, *3*(3), e1602638. https://doi.org/10.1126/sciadv.1602638

Simon, P., Pollak, L., Clevidence, B., Holden, J., & Haytowitz, D. (2009). Plant breeding for human nutrition. *Plant Breeding Reviews*, *31*, 325–392.

Simon, Philipp. (2000). Domestication, Historical Development, and Modern Breeding of Carrot. *Plant Breeding Reviews*, *19*, 157–190.

Skroch, P. W., Nienhuis, J., Beebe, S., Tohme, J., & Pedraza, F. (1998). Comparison of Mexican common bean (Phaseolus vulgaris L.) core and reserve germplasm collections. *Crop Science*, *38*(2), 488–496.

Spooner, D. M., Jansky, S. H., & Simon, R. (2009). Tests of Taxonomic and Biogeographic Predictivity: Resistance to Disease and Insect Pests in Wild Relatives of Cultivated Potato. *Crop Science*, *49*(4), 1367. https://doi.org/10.2135/cropsci2008.04.0211

Stein, M., & Nothnagel, T. (1995). Some remarks on carrot breeding (Daucus carota sativus Hoffm.). *Plant Breeding*, *114*, 1–11.

Stolarczyk, J., & Janick, J. (2011). Carrot: History and Iconography. *Chronica Horticulturae*, *51*, 13–18.

Tai, P. Y. P., & Miller, J. D. (2001). A Core Collection for Saccharum spontaneum L. from the World Collection of Sugarcane. *Crop Science*, *41*(3), 879–885. https://doi.org/10.2135/cropsci2001.413879x

Tanksley, S. D., & McCouch, S. R. (1997). Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. *Science*, *277*(5329), 1063–1066. https://doi.org/10.1126/science.277.5329.1063

Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M., & Davenport, G. F. (2009). Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. *BMC Bioinformatics*, *10*, 243. https://doi.org/10.1186/1471-2105-10-243

Theisen, T. (2016). *Organic Open Pollinated Carrot Phenotyping: Returning to the Roots for Improving Organic Production in Main and Cold Season Cultivation*. University of Wisconsin- Madison, Madison, Wisconsin.

United States Department of Agriculture Economic Research Service (USDA ERS). (2013). *Organic Production Overview: Table [data file]*. Retrieved from http://www.ars.usda.gov/

Upadhyaya, H. D., Ortiz, R., Bramel, P. J., & Singh, S. (2003). Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. *Genetic Resources and Crop Evolution*, *50*(2), 139–148. https://doi.org/10.1023/A:1022945715628

Upadhyaya, H. D., Pundir, R. P. S., Dwivedi, S. L., Gowda, C. L. L., Reddy, V. G., & Singh, S. (2009). Developing a Mini Core Collection of Sorghum for Diversified Utilization of Germplasm. *Crop Science*, *49*(5), 1769. https://doi.org/10.2135/cropsci2009.01.0014

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

What is Organic Farming? (n.d.). Retrieved January 8, 2018, from https://www.sare.org/Learning-Center/Bulletins/Transitioning-to-Organic-Production/Text-Version/What-is-Organic-Farming

Wilbois, K.-P., & Messmer, M. (2015). *Survey on the requirement in organic plant breeding*. Retrieved from http://orgprints.org/31289/1/Requirements%20in%20organic%20plant%20breeding.pdf

Willer, H., & Julia Lernoud. (n.d.). The World of Organic Agriculture Statistics and Emerging Trends 2017. Research Institute of Organic Agriculture FiBL IFOAM Organics International. Retrieved from https://shop.fibl.org/CHen/mwdownloads/download/link/id/785/?ref=1

Wright, S. (1949). The Genetical Structure of Populations. *Annals of Eugenics*, *15*(1), 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Xiurong, Z., Yingzhong, Z., Yong, C., Xiangyun, F., Qingyuan, G., Mingde, Z., & Hodgkin, T. (2000). Establishment of sesame germplasm core collection in China. *Genetic Resources and Crop Evolution*, *47*(3), 273–279. https://doi.org/10.1023/A:1008767307675

Xu, C., Gao, J., Du, Z., Li, D., Wang, Z., Li, Y., & Pang, X. (2016). Identifying the genetic diversity, genetic structure and a core collection of Ziziphus jujuba Mill. var. jujuba accessions using microsatellite markers. *Scientific Reports*, *6*, 31503. https://doi.org/10.1038/srep31503

Yildiz, M., Willis, D. K., Cavagnaro, P. F., Iorizzo, M., Abak, K., & Simon, P. W. (2013). Expression and mapping of anthocyanin biosynthesis genes in carrot. *Theoretical and Applied Genetics*, *126*(7), 1689–1702. https://doi.org/10.1007/s00122-013-2084-y

Zewdie, Y., Tong, N., & Bosland, P. (2004). Establishing a core collection of <Emphasis Type="Italic">Capsicum</Emphasis> using a cluster analysis with enlightened selection of accessions. *Genetic Resources and Crop Evolution*, *51*(2), 147–151. https://doi.org/10.1023/B:GRES.0000020858.96226.38

Zhu, Y., Chen, H., Fan, J., Wang, Y., Li, Y., Chen, J., … Mundt, C. C. (2000). Genetic diversity and disease control in rice. *Nature*, *406*(6797), 718–722. https://doi.org/10.1038/35021046

Association analysis reveals the importance of the *Or* gene in carrot (*Daucus carota* L.) carotenoid accumulation and domestication

S.L. Ellison[*,§], **C.H. Luby**[†,§], **K. Corak**[†,§], **K. Coe**[†], **D. Senalik**[*], **M. Iorizzo**[‡], **I.L. Goldman**[†], **P.W. Simon**[*] **and J.C. Dawson**[†,1]

[*]Department of Horticulture, University of Wisconsin-Madison, 1575 Linden Dr., Madison, WI 53706, [†]Agricultural Research Service, United States Department of Agriculture, 1575 Linden Dr., Madison, WI 53706, [‡]Department of Horticultural Science, North Carolina State University, 600 Laureate Way, Kannapolis, NC 28081, [§]S. Ellison, C. Luby and K. Corak contributed equally to this work.

ABSTRACT Carrots are among the richest sources of provitamin A carotenes in the human diet. Genetic variation in the carotenoid pathway does not fully explain the accumulation of such high levels of carotenoids in carrot roots. Using a diverse collection of modern and historic domesticated varieties and wild carrot accessions, an association analysis revealed a significant genomic region that contains the *Or* gene, advancing this gene as a candidate for carotenoid accumulation in carrot. Analysis of sequence variation at the *Or* locus revealed a nonsynonymous mutation co-segregating with high carotenoid content. *Or* has been found to control carotenoid accumulation in other crops but has not previously been described in carrot. Our analysis also allowed us to more completely characterize the genetic structure of carrot, showing that the Western domesticated carrot largely forms one genetic group, despite dramatic phenotypic differences among market classes. Eastern domesticated and wild accessions form a second group, which reflects the recent cultivation history of carrots in Central Asia. Other wild accessions form distinct geographic groups, with well-defined groups on the Iberian peninsula and in Northern Africa. Using genome-wide F_{st}, nucleotide diversity and XP-CLR, we analyzed the genome for regions putatively under selection during domestication, and identified twelve regions that were significant for all three methods of detection, one of which includes the *Or* gene. This provides further evidence that this gene was important in the early stages of carrot domestication and improvement and may explain why it has not been found with less genetically diverse mapping populations.

KEYWORDS *Daucus carota* | GWAS | population structure | carotenoids | domestication | selective sweep

Carrot domestication and modern breeding have been driven by selection for large roots containing abundant carotenoids, which are responsible for orange pigmentation in the taproot. The presence of carotenoids in root tissues is unlikely to confer an advantage for natural selection, but is meaningful in a domesticated context (Iorizzo *et al.* 2016) due to their visual appeal and the role of dietary pro-vitamin A compounds in human health (Arscott and Tanumihardjo 2010). Carrots are among the richest sources of provitamin A carotenes in the human diet (Simon *et al.* 2009), and significant breeding effort has focused on increasing root carotenoid accumulation (Simon 2000; Simon and Goldman 2007; Simon *et al.* 2008). Although carotenoid biosynthetic genes of carrot have been mapped (Just *et al.* 2007), they do not comprehensively explain the accumulation of high levels of carotenoids in carrot roots, leaving much of that mechanism largely unknown (Iorizzo *et al.* 2016; Ellison *et al.* 2017).

While carrot is well-known as a bright orange root crop, the original carrots domesticated in Central Asia ca. 900 CE were purple and yellow in color (Banga 1963) (Fig. 1 A,B). There is some evidence for orange carrot earlier in history (Stolarczyk and Janick 2011), but it was not until six centuries after domestication that orange roots appeared consistently in the historical record. Wild carrot is indigenous to Europe, North Africa, and Western Asia with its center of diversity in present day Afghanistan (Vavilov and Dorofeev 1992). Based on most historical records, the first evidence of carrot cultivated as a

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Friday 20th April, 2018

[1]To whom correspondence should be addressed. E-mail: dawson@hort.wisc.edu

food crop appeared in the Iranian Plateau and Persia in the 10th century (Banga 1957b,a, 1963; Brothwell and Brothwell 1969), and molecular evidence supports a Central Asian origin of domesticated carrot (Iorizzo *et al.* 2013). Carrot cultivation then spread westward to North Africa and Europe and eastward to Asia. Orange roots appeared in Spain and Germany in the 16th century (Stolarczyk and Janick 2011) and quickly became the predominant color for cultivars (Fig. 1 C,D).

Carotenoid levels have doubled due to plant breeding over the past 60 years (Simon 1990). Hence, there has been substantial effort to understand the mechanisms of carotenoid accumulation and regulation. Allelic variation at two genes, Y and Y_2, accounts for most of the distinctive color and carotenoid accumulation differences observed in orange, yellow, and white carrot roots (Buishand and Gabelman 1979). However, carotenoid biosynthesis genes in carrot do not map near enough to Y or Y_2 to be responsible for these differences (Just *et al.* 2007). The popularity of orange carrot likely fixed many of the alleles responsible for carotenoid accumulation in roots in domesticated populations. Researchers have therefore looked outside the biosynthetic pathway to regulatory and other modifying genes for explanation. Iorizzo et al. (Iorizzo *et al.* 2016) used two mapping populations and the newly assembled carrot genome to identify a candidate outside of the carotenoid biosynthetic pathway for the Y gene, *DCAR_032551*, that regulates photosystem development and conditions a portion of carotenoid accumulation in carrot roots.

In cauliflower, the *Orange* (*Or*) gene, accounts for elevated levels of carotenoid accumulation (Li *et al.* 2001). The *Or* gene is responsible for both biogenesis of chromoplasts where carotenoids are stored, and post-transcriptional regulation of Phytoene Synthase (PSY), an enzyme necessary for carotenoid biosynthesis (Yuan *et al.* 2015; Zhou *et al.* 2015; Lu *et al.* 2006). Mutations in the *Or* gene have been associated with accumulation of large amounts of carotenoids in non-leaf tissue through the differentiation of non-colored plastids into chromoplasts in arabidopsis, cauliflower, and sweet potato (Yuan *et al.* 2015). Maass et al. (Maass *et al.* 2009) noted that the accumulation of large amounts of beta-carotene in the form of crystals in carrot is strikingly similar to that found in the cauliflower *Or* mutant (Maass *et al.* 2009). Despite the accumulation of large amounts of carotenoids in orange carrot roots, the *Or* gene has not previously been associated with carotenoid accumulation in carrot. Previous carotenoid studies have focused either on biparental populations derived from crosses among domesticated carrot (Buishand and Gabelman 1979) or on crosses between wild carrot from North America and domesticated carrot (Santos and Simon 2002; Just *et al.* 2007; Ellison *et al.* 2017). Previous studies were also limited in their ability to detect significant associations by population size and marker density (Iorizzo *et al.* 2013).

In this study we genotyped 674 globally distributed domesticated and wild carrot accessions to conduct a genome wide association analysis (GWAS) for carrot root pigmentation. We also analyzed the population structure which developed during carrot dispersal and domestication. We sampled germplasm from all major global regions where carrot originated or was domesticated. Previous studies have identified three major genetic groups: Western, Eastern, and wild, but with limited numbers of accessions and low marker density (Iorizzo *et al.* 2013). Utilizing the accessions studied here, we are able to accurately represent the history of selection and breeding of the modern, domesticated orange carrot. Our analysis enabled the identification of both new and previously characterized regions of the carrot genome that were likely involved in selective sweeps during domestication and we present the first indication of the *Or* gene playing a role in carotenoid accumulation in carrot.

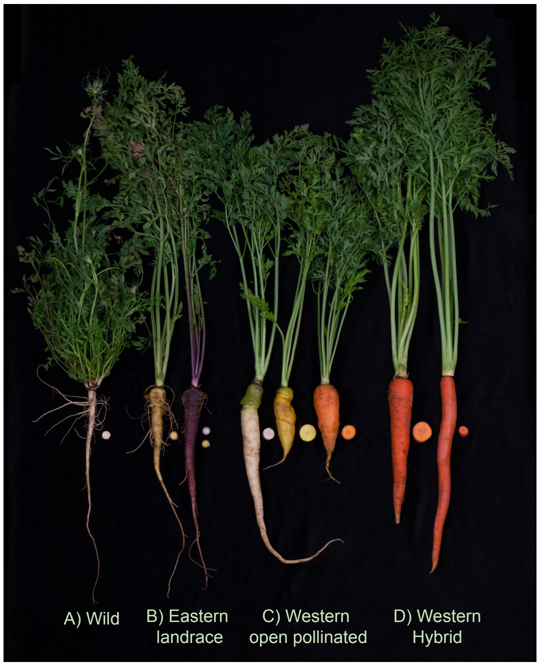


Figure 1 Carrot accessions exhibiting the range of phenotypes used in this study and the stages of carrot domestication and improvement. From L to R: (A) Wild, (B) Eastern Landrace, (C) Western Historic Open Pollinated, (D) Modern Hybrids (L: Processing type; R: Imperator type). Photo courtesy of Matthew Mirkes.

Materials and Methods

Plant Materials and Phenotypic Evaluation

Included were 705 globally distributed wild and domesticated carrot (*Daucus carota* L.) samples. Samples 1-144 were sown on certified organic land at Tipi Organic Produce in Evansville, WI, USA and Elderberry Hill Farm in Waunakee, WI, USA in the summers of 2013 and 2014. Samples 43XXX and 53XXX were grown at the West Madison Agricultural Research Station in Madison, WI, USA (WMARS) in 2014 and 2015. Samples 30XXX and 32XXX were grown at the University of Wisconsin, Hancock Agricultural Research Station in the summer of 2013 and GHXXXX, DH, and 493 samples were grown at the University of Wisconsin, Walnut Street Greenhouse in the spring of 2013. Two samples of *D. syrticus* (Ames 29096 and Ames 29108) were used as an outgroup species based on phylogenetic results of Arbizu et al. (Arbizu *et al.* 2014). Passport data for the 674 accessions can be found in Sup. Tab. S1.

Pigmentation analysis was conducted within five weeks of carrot harvest. Roots were sliced in cross section at 5-10 cm from the root tip and root phloem color was classified as orange, purple, red, white, or yellow. Visual assessment data can be found in Sup. Tab. S2. Carotenoid content was quantified using lyophilized root tissue for HPLC analysis as modified from (Simon and Wolff 1987; Simon *et al.* 1989). Briefly, 0.1 g of lyophilized carrot root tissue was crushed and then soaked in 2.0 ml of petroleum ether at 4°C. After 12-16 hours, 300 μl of the petroleum ether extract was added to 700 μl of methanol, eluted through a Rainin Microsorb-MV column and analyzed on a Millipore Waters 712 WISP HPLC system. Synthetic beta-carotene (Sigma-Aldrich, St. Louis, MO) was used in each independent run as a reference standard for calibration. Lutein, alpha-carotene, and beta-carotene were quantified by absorbance at 450 nm. Concentrations are described in μg g-1 dry weight (DW). HPLC data can be found in Sup Tab S3.

Genotyping, SNP Production and Filtering

Total genomic DNA of individual plants was isolated from approximately 0.1 g of lyophilized leaves of four-week old plants following the 10% CTAB protocol described by Murray and Thompson (Murray and Thompson 1980) with modifications by Boiteux et al. (Boiteux *et al.* 1999). All DNA was quantified using the Quantus PicoGreen dsDNA Kit (Life Technologies, Grand Island NY) and normalized to 10 ng/μl. Genotyping-by-Sequencing (GBS), as described by Elshire et al. (Elshire *et al.* 2011), was carried out at the University of Wisconsin, Madison Biotechnology Center (WI, USA) with minimal modification and half-sized reactions. Briefly, DNA samples were digested with ApeKI, barcoded and pooled for sequencing, and 80-95 pooled samples were run per single Illumina HiSeq 2000 lane, using 100 nt reads and v3 SBS reagents (Illumina, San Diego, CA). Images were analyzed using CASAVA 1.8.2. and bcl2fastq-1.8.4.

The TASSEL-GBS pipeline version 5.2.26 was used to call SNPs as described by Bradbury et al. (Bradbury *et al.* 2007) and Glaubitz et al. (Glaubitz *et al.* 2014) using the carrot reference genome (GenBank accession LNRQ01000000.1; (Iorizzo *et al.* 2016). SNPs were filtered into two datasets. D1 (Sup. Data D1) had less than 30% missing data for genotype and marker, a 5% minor allele frequency, no more than two alleles and at least 5X depth per marker. Markers were further filtered to set heterozygous markers with an allele ratio less than 0.3 or more than 0.7 to missing, leaving 39,710 SNPs in 674 genotypes. Missing genotype calls in D1 were imputed using Beagle 4.1 with niterations = 10 (Browning and Browning 2016). A subsample of D1 was created to exclude 21 wild samples from Portugal (D1-noPT). D2 (Sup. Data D2) had less than 30% missing data for genotype and 10% missing data for marker, a 5% minor allele frequency, no more than two alleles, and at least 5X depth per marker. SNPs from the resequenced outgroup samples of *D. syrticus*, Ames 29096 and Ames 29108, (Iorizzo *et al.* 2016) were added to D2 for a total of 32,128 SNPs in 676 samples. A subsample of D2 was created to exclude samples with more than 30% admixture (D2-lowAd). SNP density across chromosomes, using 500,000 nt bins for D1 and D2 can be found in Sup. Figs. S1 and S2. Filtering parameters for each SNP dataset can be found in Sup. Fig. S3. SNP datasets are in Sup. Data D1 and Sup. Data D2.

Linkage Disequilibrium

TASSEL 5 (Bradbury *et al.* 2007) was used to calculate LD for the full matrix of SNPs for dataset D1-noPT. Reported values of LD decay use an r^2 cutoff of 0.1 and 0.2 for filtered SNPs ($p < 0.01$) (Vos *et al.* 2017). The half distance of LD decay was calculated as when the LD decay curve intersects with half the maximum LD value. Genome-wide sliding window analysis of LD was conducted for both wild and domesticated samples using VCFtools with the parameters –geno-r2 –ldwindow 100 (51). r^2 values with fewer than 95 SNPs per bin were removed. Sliding window analysis was visualized using qqman in R studio (Wickham 2009).

Population Structure

We used Dataset D2 and conducted eight replications of the Bayesian clustering program STRUCTURE version 2.3.4 (Pritchard *et al.* 2000) with populations (K value) ranging from 1 to 14, with a burn-in length of 20,000 and 50,000 Monte Carlo iterations, respectively. An admixture model with no previous population information was included; all other parameters were set to default values. STRUCTURE results were processed in the software STRUCTURE HARVESTER 0.6.94 with parameter –evanno (Earl and von Holdt 2012) to detect the most likely number of clusters by using the rate of change in the log probability between successive values of K (ΔK) (Evanno *et al.* 2005). Population structure was visualized using distruct software version 1.1 (Rosenberg 2004).

Principal Component Analysis

An eigenvalue decomposition of the SNP covariance matrix was performed using TASSEL 5 using default parameters for D2 and D2-lowAd. All individuals' loadings were plotted along the first and second principal components using ggplot in R. Individuals were colored according to their STRUCTURE group identity.

Maximum-likelihood tree (RAxML)

Using Datasets D2 and D2-lowAd, maximum likelihood analyses were conducted with the GTR+G nucleotide substitution model using RAxML version 8.2.9 (Stamatakis 2014). GATK HaplotypeCaller (McKenna *et al.* 2010) with parameters –genotyping_mode GENOTYPE_GIVEN_ALLELES was used to call SNPs for the two outgroup accessions, *D. syrticus*, SRR2147152 and SRR2147153 (Arbizu *et al.* 2016). FigTree (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize phylogenetic trees.

Pairwise F_{st}

Weir and Cockerham's method for calculating pairwise F_{st} (Weir and Cockerham 1984) was implemented within the genet.dist function of the R package hierfstat (Goudet 2005). Pairwise values were calculated on all K=6 subpopulations using Datasets D2 and D2-lowAd. The dataset was first converted to a genind object using the df2genind command of the R package adegenet using default parameters.

Sliding Window Analysis of Nucleotide Diversity, F_{st}, and XP-CLR

Selective Sweep detection analyses used Dataset D1-noPT. VCFtools was used to calculate genetic diversity (π) in 500 kb windows across the carrot genome (–window-pi 500000) for wild and domesticated carrot samples. Potential selective sweep regions were found by calculating the difference between wild and domesticated nucleotide diversity bins and selecting bins in the top 5% of values ($\pi > 1.578$). The population differentiation statistic, F_{st} was estimated between wild and domesticated samples in VCFtools in 500 kb windows with 100 kb steps (-weir -fst-pop -fst-window-size 500000-fst-window-step 100000) (Danecek *et al.* 2011). Potential sweep regions were defined as the top 5% of values that were calculated ($F_{st} > 0.29$). A third method, XP-CLR, was implemented to test for selective sweeps (Chen *et al.* 2010). The XP-CLR software was run with parameters: -w1 0.005 50 100 1 -p1 0.9 for each chromosome. The genetic distances between SNPs were interpolated according to their physical distances in a high-density genetic map from the carrot genome manuscript (Iorizzo *et al.* 2016). Mean XP-CLR scores were tabulated in non-overlapping 10 kb windows across the genome. Windows with the top 1% of XP-CLR values (11.93) were selected and placed in corresponding bins from the F_{st} and nucleotide diversity analyses. Genome-wide sliding window analyses were plotted using the R package qqman (Turner 2014). Overlapping genomic regions in the top 5% for nucleotide diversity and F_{st} and top 1% XP-CLR scores were presented in a Venn diagram to uncover the most likely selective sweeps.

Genome-Wide Association Analysis

A genome-wide association analysis was performed for carrot root pigmentation using Dataset D1 by implementing the EGSCORE function in the GenABLE R package (Aulchenko *et al.* 2007). The following parameters were used: naxes=2, times=1, quiet=FALSE, bcast=0, clamda=T, propsPs=1. No fixed effects were included as covariates. The kinship matrix was calculated using the ibs command in GenABLE with the weight parameter set to “freq”. The diagonal of the kinship matrix was replaced with the variance of the average homozygosity within each individual. Manhattan and qqplots were drawn using the R package qqman (Turner 2014).

Observed Heterozygosity (H_o) and Gene Diversity (H_s)

Observed heterozygosity H_o, within population gene diversity (H_s), overall gene diversity (H_t) and overall F_{st} were calculated using the basic.stats function in the R package hierfstat (Goudet 2005) using Datasets D1, D2 and D2 lowAd. Datasets were first converted to genind objects using the df2genind command of the R package adegenet using default parameters.

Candidate Gene Sequence Analysis

Thirteen previously resequenced carrot PIs (Sup. Tab. S4) were surveyed for any sequence variation within the open reading frame of the *Or* gene (DCAR_009172). One SNP was identified between low and high carotenoid genotypes within exon 5. A transition of T to C at position 3350 resulted in a change of the codon TTG to TCG, causing a missense mutation of Leucine to Serine. This SNP is located on chromosome 3, position 5197361. In order to genotype carrot PIs for T3350C, primers that flank the SNP were generated (Sup. Tab. S5). PCR based sequencing was performed on 197 domesticated and 82 wild carrot PIs. Sequencing results were analyzed using sequencer. A gene model for *Or* was generated from the website http://wormweb.org/exonintron. Phenotypic differences for lutein, alpha-carotene, and beta-carotene were analyzed for the three *Or* genotypic classes. For each trait, significance between different genotypic classes was determined by using the aov and TukeyHSD functions in R.

Carrot sequences used for the *Or* gene alignment and *D. syrticus* samples used as an outgroup for phylogenetic analysis are available under the National Center for Biotechnology Information (NCBI) Bioproject accession PRJNA291976. All data sets necessary to reproduce the analyses and figures in this manuscript are available on FigShare.

Results

SNP Discovery

Two datasets comprising 154 wild and 520 domesticated carrots (Fig. 2 B,E, Sup. Tab. S1) were genotyped to maximize geographic distribution and minimize ascertainment bias. After filtering for missing data (< 0.3), minor allele frequency (< 0.05), coverage ($> 5\times$), allele count (≤ 2) and imputing missing data, Dataset D1 (Sup. Data D1) had a total of 39,710 SNPs in 674 individuals. The average SNP distribution across the carrot genome was approximately 54 SNPs per 500 kb bin or $\sim$ 1 SNP per 10 kb (Sup. Fig. S1) with an average 18× coverage per SNP. The same filtering parameters were used for Dataset D2 (Sup. Data D2) except SNPs were filtered using 10% missing data and were not imputed. Additionally two samples from the outgroup *Daucus syrticus* were included for a total of 676 individuals and 32,128 SNPs. SNP distribution for D2 was similar to D1 with 43 SNPs per 500kb (Sup. Fig. S2) and 20X coverage per SNP. Additional information about SNP filtering can be found in Sup Fig. S3.

Rapid Decay of Linkage Disequilibrium in Carrot

LD analysis of wild carrot accessions demonstrated a very rapid genome-wide decay between $\sim$ 100 bp ($r^2 = 0.2$) and $\sim$ 1 kb ($r^2 = 0.1$) and a rapid decay of $\sim$ 400 bp ($r^2 = 0.2$) and $\sim$ 13 kb ($r^2 = 0.1$) in domesticated accessions. This rapid decay was further supported by estimates of wild and domesticated samples having an LD half life of 67 bp and 6,544 bp, respectively (Sup. Fig. S4). Determination of LD decay distances does not have a consensus method in the literature, with both thresholds (0.1 and 0.2) and half-life methods used (Vos *et al.* 2017). Half life methods may be more robust to differences in minor allele frequencies and have been used in a number of species (Vos *et al.* 2017; Branca *et al.* 2011; Kim *et al.* 2007; Lam *et al.* 2010; Zhao *et al.* 2011).

The pattern of LD in a genome is a powerful signal of the population genetic processes that are structuring it, and similar LD decay rates have been found in other highly heterozygous outcrossing species such as maize and grape (Yan *et al.* 2009; Myles *et al.* 2011). The observed rapid decay suggests genome-wide association studies should be very useful for identifying candidate genes in carrot as long as SNP density and coverage is comprehensive.

Population Structure Dynamics among Wild and Domesticated Carrot

Selection by humans has resulted in phenotypic differences between domesticated and wild carrots for traits such as flavor, biennial growth habit, root system architecture, disease resistance, and root pigmentation (Simon 2000). In addition to being phenotypically distinguishable, previous studies have demonstrated that wild and domesticated carrots are genetically distinct (Iorizzo *et al.* 2013; Baranski *et al.* 2012; Clotault *et al.* 2010; Shim and Jorgensen 2000; Rong *et al.* 2014) and also that they separate into geographically discrete Eastern and Western groups (Baranski *et al.* 2012; Iorizzo *et al.* 2013; Grzebelus *et al.* 2014; Iorizzo *et al.* 2016).

An examination of population structure was carried out using STRUCTURE software with K=6 as the number of groups strongly supported by the Evanno method (Evanno *et al.* 2005) (Fig. 2 A, Sup. Figs. S5, S6, S7, S8). The support for K=6 was slightly stronger than K=4 or K=5, and as we are interested in understanding population structure in carrot we chose to work with the largest K value strongly supported by the data.

To maximize cluster separation, a low admixture group (D2-LowAd) of 463 accessions was created by only including samples when the proportion of inferred ancestry was greater than seventy percent ($q \geqslant 70\%$)(Fig. 2 E, Sup. Tab. S1). Clustering with STRUCTURE indicated divisions between Western domesticated (-D), Western wild (-W), and all Eastern (-D/W) samples, as well as emergent subclusters corresponding to geographic origin including wild samples from Tunisia (-W) and wild accessions from Portugal (-W) (Fig. 2 A). An additional cluster formed for Western Imperator hybrids (Western-HI) (Fig. 1 D, Fig. 2 A). The Q matrix of individual accessions is reported in Sup. Tab. S1.

The observed population substructure was supported by phylogenetic analysis, PCA, and pairwise F_{st}. Using *D. syrticus* as an outgroup (Arbizu *et al.* 2014), the maximum-likelihood analysis identified the same six strongly supported clades (bootstrap $> 97\%$): Portuguese-W, Western-W, Eastern-W/D, Tunisian-W, Western-D, and Western-HI carrots (Fig. 2 D). PCA revealed a clear separation between wild and domesticated carrots along the first principal component (12.4% of variation explained) and between Eastern and Western samples along the second principal component (4.6% of variation explained, Fig. 2 C). Pairwise F_{st} calculations further supported differentiation between the six subclusters (Sup. Tab. S7). The Portuguese-W samples were the most strongly divergent of all the STRUCTURE groups, forming a very distinct subpopulation separate from other wild carrot accessions. Observed heterozygosity (H_o) for accessions in dataset D2-lowAd was 0.18 (Sup. Tab. S8). All analyses and results were also confirmed on Dataset D2, without removal of high admixture samples (Sup. Fig. S6 and Sup. Tab S7, Sup. Tab. S8).

GWAS Analysis Identifies Or as a Candidate Gene for Carotenoid Accumulation in Carrot

To identify genomic regions potentially related to carotenoid accumulation, we performed a GWAS for orange pigmentation in carrot root using Dataset D1 (Sup. Fig. S3). We found a previously unidentified significant 143 kb GWAS signal on chromosome 3 containing 17 annotated genes (Fig. 3 A, Sup. Tab. S9). Inspection of the Q-Q plot demonstrated an inflated observed p-value (Sup. Fig. S9) that is likely attributed to the orange phenotype being absent in wild carrot accessions but common in domesticated accessions, causing the effects of population and polymorphisms to be confounded (Korte and Farlow 2013). *Or*, a gene associated with carotenoid biosynthesis regulation and chromoplast formation (Zhou *et al.* 2015; Li *et al.* 2012; Lu *et al.* 2006), is in the middle of the 143 kb region encompassing the most significant SNPs in our GWAS analysis. No other genes in the 143 kb region are known to be associated with carotenoid accumulation. To better characterize the association of carotenoid accumulation and the *Or* gene we looked for mutations co-segregating between five high and eight low carotenoid accessions that had been previously resequenced (Iorizzo *et al.* 2016) and found a nonsynonmous mutation at position 3350 in exon 5, causing a Serine to Leucine amino acid change (Fig. 3 B). An additional 198 domesticated samples were phenotyped for lutein, alpha-carotene, and beta-carotene content using HPLC and genotyped at *Or*. Those samples with the T/T genotype had significantly higher amounts of alpha- and beta-carotene then those heterozygous (C/T) or homozygous recessive (C/C) (Fig. 3 C, Sup. Tab. S3). The same was true for lutein, however, the heterozygous group could not be significantly differentiated from either homozygous group. Eighty-two wild carrot (low-carotenoid) samples were genotyped at *Or* and all samples had the low carotenoid C/C genotype. This is the first report of an association between *Or* and carotenoid accumulation in carrot.

Identification of Selection Signatures during Carrot Domestication

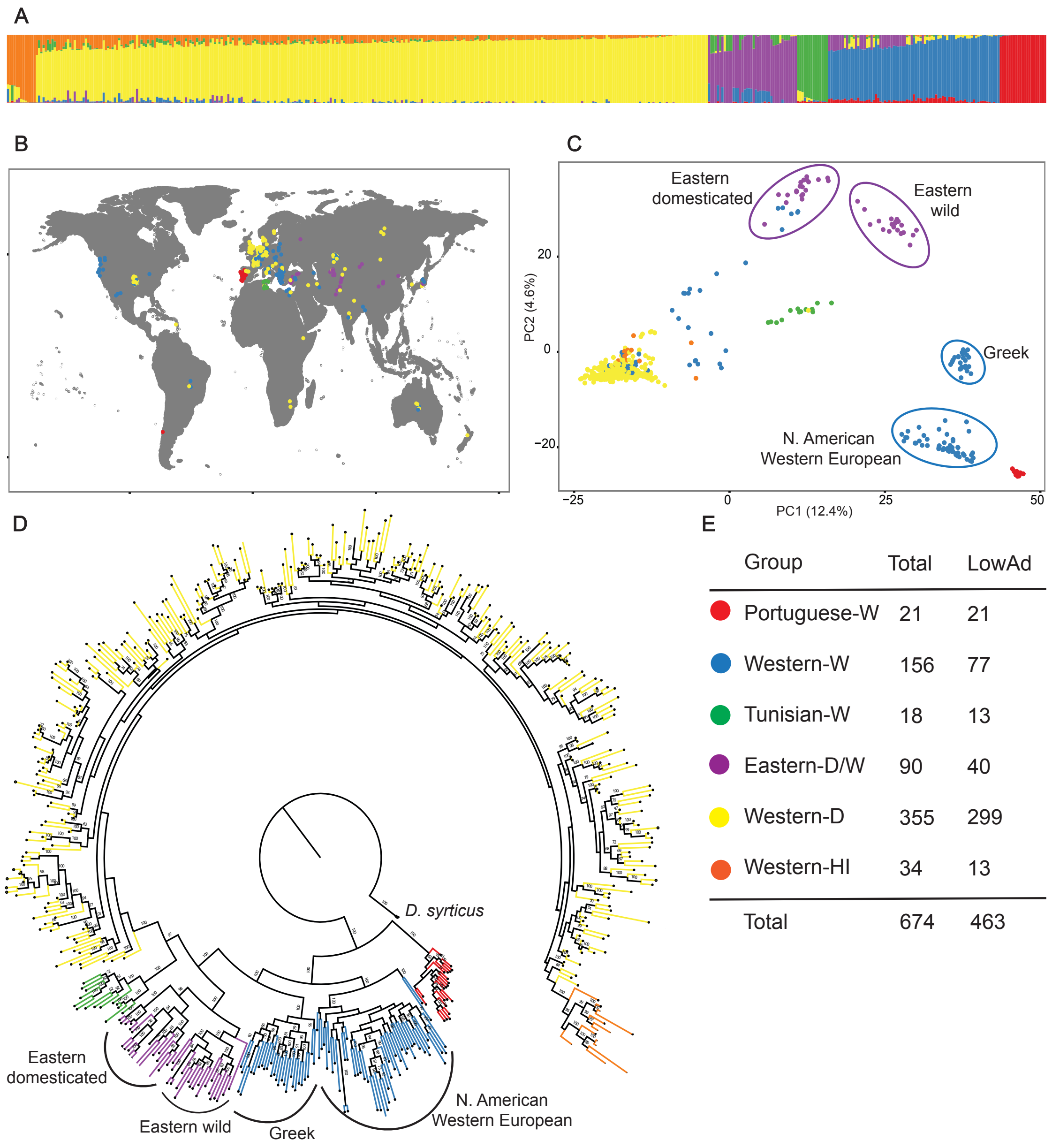
During crop domestication highly favorable alleles undergo intensive selection and reach fixation rapidly resulting in reduced variation in neighboring genomic regions thereby creating a signature of a selective sweep. We used three measures to analyze sweeps: reduced nucleotide diversity (π) (Nei and Li 1979) in domesticated samples as compared to wild, high population differentiation (F_{st}) (Wright 1951) between wild and domesticated samples, and allele frequency differentiation between populations (XP-CLR) (Chen *et al.* 2010). To reduce potential confounding effects of population structure and differentiation we removed the 21 Portuguese-W samples from the selective sweep analyses (Dataset D1-noPT, Sup. Fig. S3). Differences in nucleotide diversity between wild and domesticated samples were estimated for 500 kb bins across the carrot genome. The average difference between groups was 1.080 with 37 potential selective sweep regions detected using the top 5% of calculated values (1.578) (Fig. 4 A and Sup. Tab. S6).

Overall, we found little reduction in genetic diversity in all domesticated carrot (3.13×10^{-5}) compared to all wild carrot (3.25×10^{-5}), averaged across the whole genome level.

The genome-wide average F_{st} between domesticated and wild carrot was 0.14. We detected 38 genomic regions with F_{st} values above the 95% percentile (> 0.29), differentiating wild and domesticated accessions (Fig. 4 A and Sup. Tab. S6). These regions with high levels of differentiation likely experienced selective sweeps during domestication or improvement (Wright 1951). The recently identified *Y* gene (Iorizzo *et al.* 2016), a candidate for carotenoid accumulation in carrot taproot is located within one of these regions of high differentiation between wild and domesticated carrots (24.5-25.0 Mb on chromosome 5). The carotene hydroxylase *DcCYP97A3* gene associated with increased alpha-carotene maps near another region of high differentiation on chromosome 7 (6.5-7.0 Mb) (Arango *et al.* 2014 Carotene hydroxylase).

Lastly we used the cross-population composite likelihood ratio (XP-CLR) method to compare the wild and domesticated accessions in 10kb bins across the genome (Chen *et al.* 2010). The top 1% of XP-CLR values (> 11.94), identified 78 potential sweeps bins (Fig. 4 A and Sup. Tab. S6). A candidate domestication gene associated with root-thickening, *DcAHLc1* (Macko-Podgorni *et al.* 2017), is located at 41.8Mb on chromosome 2, near one of the regions with the highest XP-CLR scores (42.0-42.5 Mb). Another region, 33.5-34.0 Mb on chromosome 7, overlaps with the recently fine-mapped QTL, *Y2*, a gene associated with carotenoid accumulation (Ellison *et al.* 2017).

To identify the most supported potential selective sweeps during domestication, we considered regions that were significant for all three methods of detection used (decreased nucleotide diversity, increased F_{st}, and a high XP-CLR score). Using that approach, 12 such regions were identified in comparing wild and domesticated carrot accessions (Fig. 4 A and B, Sup. Figs. S10, S11, S12). The candidate carotenoid accumulation gene, *Or*, which was identified in our GWAS falls in one of these 12 genomic locations. A genome-wide sliding window analysis of LD also identified the same region on chromosome 3 to have the slowest LD decay in domesticated carrots (Fig. 4 C) but not wild carrots (Sup. Fig. S13). These results strongly suggest that selection pressures acted on the *Or* locus during carrot domestication. It is possible that high-carotene alleles at the *Or* locus have been fixed in most western domesticated carrots, which may explain



Group	Total	LowAd
● Portuguese-W	21	21
● Western-W	156	77
● Tunisian-W	18	13
● Eastern-D/W	90	40
● Western-D	355	299
● Western-HI	34	13
Total	674	463

Figure 2 Population structure of 463 carrot accessions with $<$ 30% admixture (D2-LowAd). A) STRUCTURE groups. Percentage of membership (q) for each group identified at K=6. B) Geographic distribution of accessions each represented by a point on the map colored according to STRUCTURE group. Current commercial varieties not shown. C) PCA plot of the first two principal components. PC1 and PC2 account for 12.4% and 4.6% of the total variation, respectively. D) Maximum-likelihood tree of carrot accessions. Numbers on the branches indicate bootstrap support. Black branch represents outgroup *D. syrticus*. E) Color key. Total number of accessions in each STRUCTURE Group.

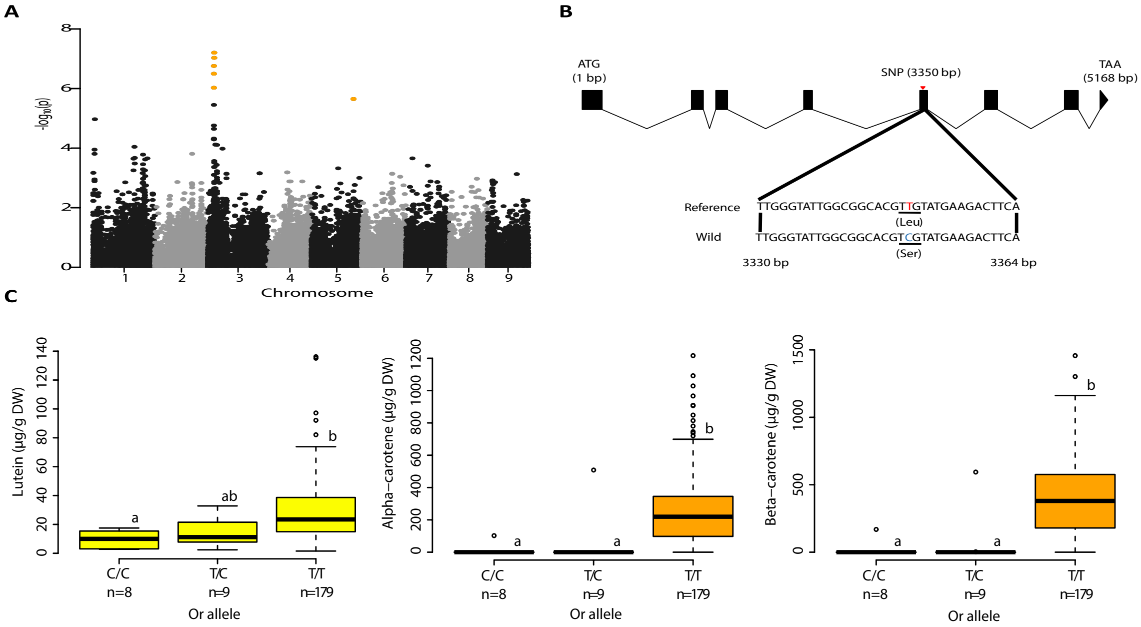


Figure 3 Genome-wide association analysis of orange pigmentation and identification of the candidate gene *Or* on chromosome 3. A) Manhattan plot for orange carrot root color. Orange SNPs, with empirically-adjusted p-values less than 0.05, were defined as significant. B) Open reading frame of *Or* and the nonsynonymous mutation in exon 5 at position 3350 (T3350C). C) Box plots for lutein, alpha-carotene, and beta-carotene for the three *Or* genotypes (C/C, T/C, and TT) at position 3350. Center line = median, box limits = upper and lower quartiles, whiskers = 1.5× the interquartile range, dots = outliers. Different letters indicate significant differences between genotypes (P <0.05, Tukey's HSD).

why it was not identified until a globally distributed dataset of wild and domesticated carrots was used.

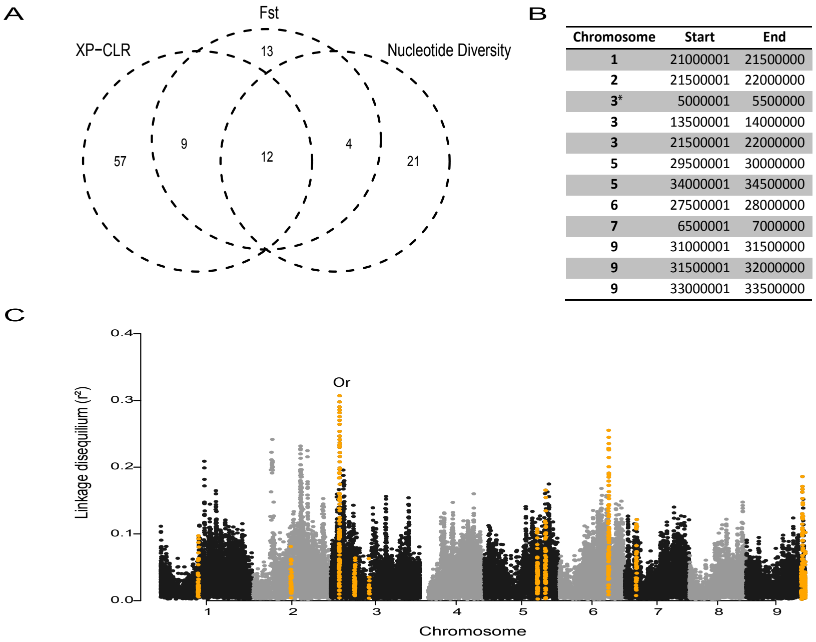
Discussion

In this study, we genotyped a large and diverse collection of carrot accessions to determine the global structure of LD in the genome. Genome-wide coverage was approximately 1 SNP per 10 kb, dense enough to give an initial assessment of the pattern of LD in carrot. We find LD decays very rapidly in both wild and domesticated accessions with a half life of 67 bp and 6,544 bp, respectively (Sup. Fig. S4) and we also demonstrate that LD decline is variable across the nine chromosomes as well as between wild and domesticated accessions (Fig. 4 C, Sup. Fig. S13). Future GWAS and LD projects will benefit from improved genotyping techniques, such as resequencing or two-enzyme GBS (Poland *et al.* 2012), to increase SNP density across the genome.

The primary divisions of population structure across our diverse carrot accessions are geographic distribution, west to east, and intensity of breeding effort, wild to domesticated. As previously demonstrated, most variation occurs between wild and domesticated accessions (Iorizzo *et al.* 2013; Grzebelus *et al.* 2014; Rong *et al.* 2014), however, there is evidence of continued gene flow where populations overlap geographically, such as in Western-W accessions which are present in areas where domesticated carrot is grown. It also appears that there is significant overlap in wild and domesticated samples from the Eastern group. This may be attributed to either recent admixture or to domesticated carrots sharing many of the same alleles as wild carrots from the region. While STRUCTURE failed to identify a distinction between Eastern wild and Eastern domesticated carrots, these do appear as sister clades in the phylogeny with wild Western carrots at the root of both clades (Fig. 2 D), supporting recent findings that domesticated carrots are genetically closer to Eastern wild carrots than to Western wild carrots (Iorizzo *et al.* 2013; Vavilov and Dorofeev 1992).

Carrots from Northern Africa, Tunisian-W, form a distinct group but show the least differentiation from all other groups (Sup. Tab. S7). Previously North African samples clustered closer to wild samples from the West and Middle East (Iorizzo *et al.* 2013) but here, using a much larger dataset and number of SNPs, the maximum-likelihood analysis places Tunisian-W samples at the base of all domesticated western carrots (Fig. 2 D), suggesting carrots from this region of the world may have been important for the improvement of domesticated carrots. Future field sampling efforts and population dynamics analysis should include more representation from North Africa to better understand carrot domestication and diversity. Finally we observe Portuguese-W samples are highly diverged from other accessions. Gene flow in and out of the Iberian peninsula region is likely limited because of the Pyrenees mountain range. However, crosses with Western domesticated carrot have been successful and therefore Portuguese-W samples may provide a novel source of alleles for abiotic stresses.

The analysis of an extensive and representative sample of modern domesticated, historic domesticated, and wild accessions allowed us to identify genomic regions putatively under selection. False positives can be exacerbated by large genomic datasets so we used a conservative approach to only consider regions identified by all three detection tests (decreased nucleotide diversity, high F_{st}, and elevated XP-CLR scores) and identified 12 putative genomic regions under selection during domestica-



Chromosome	Start	End
1	21000001	21500000
2	21500001	22000000
3*	5000001	5500000
3	13500001	14000000
3	21500001	22000000
5	29500001	30000000
5	34000001	34500000
6	27500001	28000000
7	6500001	7000000
9	31000001	31500000
9	31500001	32000000
9	33000001	33500000

Figure 4 Regions of the carrot genome that likely underwent a selective sweep during domestication. A) Venn diagram represents the overlapping of 500 kb regions tested for selection signatures - top 5% of F_{st} and nucleotide diversity difference between wild and domesticated carrot accessions and top 1% of XP-CLR values. B) Genomic location of potential selective sweeps identified by F_{st}, nucleotide diversity and XP-CLR. The asterisk signifies the genome region carrying the candidate orange pigmentation gene, *Or*. C) Genome-wide linkage disequilibrium averaged across sliding windows of 100 SNPs in domesticated carrots. Regions identified as significant in A and B are highlighted in orange. The region containing the *Or* candidate gene for orange pigmentation in carrot is marked 'Or'.

tion (Fig. 4 A,B). One selective sweep located on chromosome 3 overlapped with the most significant SNPs in our GWAS analysis for carotenoid accumulation and contained the candidate gene *Or*. Analysis of the *Or* sequence between samples with varying carotenoid content found a nonsynonymous mutation in exon 5 that associates with increased quantities of alpha- and beta-carotene and to a lesser extent lutein. Single amino acid substitutions in the Or homologs in melon and Arabidopsis have lead to increase carotenoid accumulation (Tzuri *et al.* 2015; Yuan *et al.* 2015).

Or is important for chromoplast development, a necessary precursor to carotenoid accumulation (Lu *et al.* 2006). *Or* differentiates non-colored plastids into chromoplasts, which provide the deposition sink for carotenoid accumulation (Lu *et al.* 2006). *Or* also post-transcriptionally regulates Phytoene Synthase (PSY), the most important regulatory enzyme in the carotenoid pathway (Zhou *et al.* 2015; Li *et al.* 2012; Park *et al.* 2016). This post-transcriptional effect may be why *Or* has not been identified in previous carrot studies that have looked at carotenoid accumulation mechanisms at the transcription level (Simpson *et al.* 2016). Mutations in the *Or* gene are associated with increased chromoplast formation thereby providing more storage capability for carotenoid accumulation (Yuan *et al.* 2015). We hypothesize that a mutation in *Or* enhanced carotenoid sequestration by optimizing chromoplast formation and likely was selected in conjunction with or predated carotenoid accumulation mutations such as y and y_2, during carrot domestication.

This study brings us one step closer to understanding how carrots accumulate significant levels of carotenoids. Future work should analyze *Or* expression at the transcript and protein levels, and verify the effect of disrupting its functionality on carotenoid accumulation. Additionally, the 11 other genomic regions showing consistent signatures of selection (Fig. 4 A,B) should be explored for candidate domestication genes and be considered in tandem with GWAS and mapping studies. Understanding the genetic consequences of domestication and selection on carrot can inform future plant breeding efforts and allow us to achieve greater gains from selection.

Acknowledgements The authors thank Rob Kane (deceased), Paul Miller, Brian Emerson, Thomas Hickey, Steve Pincus, Beth Kazmar, Eric Elderbrock and Holly Ruess for field assistance and Terri Theisen for phenotyping historic accessions. We also wish to thank Dr. David Spooner for technical comments to the manuscript. The authors also thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility for providing genotyping-by-sequencing library preparation and sequencing services. S.E. was supported by the NSF under grant 1202666 and the National Institute of Food and Agriculture, USDA (NIFA-USDA), under award number 2016-51181-25400. M.I. was supported by NIFA-USDA Hatch project 1008691 and K.C. was supported by NIFA-USDA Hatch project 1002863 and NIFA-USDA award number 2017-67013-26187. C.L. was supported by The Clif Bar Family Foundation Seed Matters Fellowship and NIFA-USDA North Central Region-SARE graduate student grant (GNC13-175). KCoe was supported by a University of Wisconsin-Madison Science and Medicine Graduate Research Scholars Fellowship. We also thank the many seed companies and plant breeders that provided seed for this project.

Literature Cited

Arango, J., M. Jourdan, E. Geoffriau, P. Beyer, and R. Welsch, 2014 Carotene hydroxylase activity determines the levels of both alpha-carotene and total carotenoids in orange carrots. *Plant Cell* **26**: 2223–2233.

Arbizu, C., S. Ellison, D. Senalik, P. Simon, and D. Spooner, 2016 Genotyping-by-sequencing provides the discriminating power to investigate the subspecies of *Daucus carota* (Apiaceae). *BMC Evolutionary Biology* **16**: 234.

Arbizu, C., H. Ruess, D. Senalik, P. Simon, and D. Spooner, 2014 Phylogenomics of the carrot genus (*Daucus*, Apiaceae). *American Journal of Botany* **101**: 1666–1685.

Arscott, S. and S. Tanumihardjo, 2010 Carrots of many colors provide basic nutrition and bioavailable phytochemicals acting as a functional food. *Comprehensive reviews in food science and food safety* **9**: 223–239.

Aulchenko, Y., S. Ripke, A. Isaacs, and C. van Duijn, 2007 GenABEL: an R package for genome-wide association analysis. *Bioinformatics* **23**: 1294–1296.

Banga, O., 1957a The development of the original European carrot material. *Euphytica* **6**: 64–76.

Banga, O., 1957b Origin of the European cultivated carrot. *Euphytica* **6**: 54–63.

Banga, O., 1963 Main Types of the Western Carotene Carrot and Their Origin. (W.E.J. Tjeenk Willink, Zwolle [Netherlands]). OCLC: 464970.

Baranski, R. *et al.*, 2012 Genetic diversity of carrot (*Daucus carota* L.) cultivars revealed by analysis of SSR loci. *Genet Resour Crop Ev.* **59**: 163–170.

Boiteux, L., M. Fonseca, and P. Simon, 1999 Effects of plant tissue and DNA purification method on randomly amplified polymorphic DNA-based genetic fingerprinting analysis in carrot. *J. Am. Soc. Hortic. Sci.* **124**: 32–38.

Bradbury, P., Z. Zhang, D. Kroon, T. Casstevens, and Y. Ramdoss, 2007 TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.

Branca, A. *et al.*, 2011 Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume medicago truncatula. *Proceedings of the National Academy of Sciences* **108**: E864–70.

Brothwell, D. and P. Brothwell, 1969 Food in Antiquity: A Survey of the Diet of Early Peoples. (Thames & Hudson,, London, UK).

Browning, B. and S. Browning, 2016 Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**: 116–126.

Buishand, J. and W. Gabelman, 1979 Investigations on the inheritance of color and carotenoid content in phloem and xylem of carrot roots (*Daucus carota* L.). *Euphytica* **28**: 611–632.

Chen, H., N. Patterson, , and D. Reich, 2010 Population differentiation as a test for selective sweeps. *Genome Research* **20**: 393–402.

Clotault, J., E. Geoffriau, E. Lionneton, M. Briard, and D. Peltier, 2010 Carotenoid biosynthesis genes provide evidence of geographical subdivision and extensive linkage disequilibrium in the carrot. *Theor Appl Genet* **121**: 659–672.

Danecek, P. *et al.*, 2011 The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Earl, D. and B. von Holdt, 2012 STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genet Resour.* **4**: 359–361.

Ellison, S., D. Senalik, H. Bostan, M. Iorizzo, and P. Simon, 2017 Fine mapping, transcriptome analysis, and marker development for Y2, the gene that conditions β-carotene accumulation in carrot (*Daucus carota* L). *G3: Genes, Genomes, Genetics* **7**: 2665–2675.

Elshire, R. *et al.*, 2011 A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLOS One* .

Evanno, G., S. Regnaut, and J. Gouder, 2005 Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol.* **14**: 2611–2620.

Glaubitz, J. *et al.*, 2014 TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE **9**: e90346.

Goudet, J., 2005 Hierfstat, a package for r to compute and test hierarchical f-statistics. *Molecular Ecology Notes* **5**: 184–186.

Grzebelus, D. *et al.*, 2014 Diversity, genetic mapping, and signatures of domestication in the carrot (*Daucus carota* L.) genome, as revealed by Diversity Arrays Technology (DArT) markers. *Molecular Breeding* **33**: 625–637.

Iorizzo, M. *et al.*, 2013 Genetic structure and domestication of carrot (*Daucus carota* subsp. *sativus*) (Apiaceae). *American Journal of Botany* **100**: 930–938.

Iorizzo, M. *et al.*, 2016 A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics* **48**: 657–666.

Just, B. *et al.*, 2007 Carotenoid biosynthesis structural genes in carrot (*Daucus carota*): Isolation, sequence-characterization, single nucleotide polymorphism (SNP) markers and genome mapping. *Theor Appl Genet.* **114**: 693–704.

Kim, S. *et al.*, 2007 Recombination and linkage disequilibrium in arabidopsis thaliana. *Nature Genetics* **39**: 1151–1155.

Korte, A. and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* **9**: 29.

Lam, H. *et al.*, 2010 Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* **42**: 1053–1059.

Li, L. *et al.*, 2012 The *Or* gene enhances carotenoid accumulation and stability during post-harvest storage of potato tubers. *Molecular Plant* **5**: 339–352.

Li, L., D. Paolillo, M. Parthasarathy, E. Dimuzio, and D. Garvin, 2001 A novel gene mutation that confers abnormal patterns of beta-carotene accumulation in cauliflower (*Brassica oleracea* var. *Botrytis*). *The Plant Journal: For Cell and Molecular Biology* **26**: 59–67.

Lu, S. *et al.*, 2006 The cauliflower *Or* gene encodes a DnaJ cysteine-rich domain-containing protein that mediates high levels of β-carotene accumulation. *The Plant Cell Online* **18**: 3594–3605.

Maass, D., J. Arango, F. Wüst, P. Beyer, and R. Welsch, 2009 Carotenoid crystal formation in arabidopsis and carrot roots caused by increased phytoene synthase protein levels. *PLOS ONE* **4**: e6373.

Macko-Podgorni, A. *et al.*, 2017 Characterization of a genomic region under selection in cultivated carrot (*Daucus carota* subsp. *sativus* reveals a candidate domestication gene. *Frontiers in Plant Science* **8**.

McKenna, A. *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Murray, M. and W. Thompson, 1980 Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**: 4321–4326.

Myles, S. *et al.*, 2011 Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences* **108**: 3530–3535.

Nei, M. and W. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. *PNAS* **76**: 5269–5273.

Park, S. *et al.*, 2016 Orange protein has a role in phytoene synthase stabilization in sweet potato. *Scientific Reports* **6**: 33563.

Poland, J., P. Brown, M. Sorrells, and J. Jannink, 2012 Development of high-densit genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLOS ONE* **7**: e32253.

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

Rong, J. *et al.*, 2014 New insights into domestication of carrot from root transcriptome analyses. *BMC Genomics* **15**: 895.

Rosenberg, N., 2004 DISTRUCT: A program for the graphical display of population structure. *Mol Ecol Notes* **4**: 137–138.

Santos, C. and P. Simon, 2002 QTL analyses reveal clustered loci for accumulation of major provitamin A carotenes and lycopene in carrot roots. *Molecular genetics and genomics: MGG* **268**: 122–9.

Shim, S. and R. Jorgensen, 2000 Genetic structure in cultivated and wild carrots (*Daucus carota* L.) revealed by AFLP analysis. *Theor Appl Genet.* **101**: 227–233.

Simon, P., 1990 Carrots and other horticultural crops as a source of provitamin a carotenes. *HortScience* **25**: 1495–1499.

Simon, P., 2000 Domestication, historical development, and modern breeding of carrot. *Plant Breeding Rev* **19**: 157–190.

Simon, P. *et al.*, 2008 Carrot. In Handbook of Crop Breeding, Volume 1,Vegetable Breeding, edited by J. P. MC and N. F., pp. 327–357, Springer-Verlag, GmBH, Heidelberg, Germany.

Simon, P. and I. Goldman, 2007 Carrot. In Genetic Resources, Chromosome Engineering and Crop Improvement: Vegetable Crops., edited by S. RJ, (CRC Press, Boca Raton).

Simon, P., L. Pollak, B. Clevidence, J. Holden, and D. Haytowitz, 2009 Plant breeding for human nutrition. *Plant Breeding Reviews* **31**: 325–392.

Simon, P. and X. Wolff, 1987 Carotenes in typical and dark orange carrots. J. Agric. Food Chem. **35**: 1017–1022.

Simon, P., X. Wolff, C. Peterson, , and D. Kammerlohr, 1989 High carotene mass carrot population. *HortScience* **24**: 174–175.

Simpson, K., A. Cerda, and C. Stange, 2016 *Carotenoid biosynthesis in* Daucus carota. *SpringerLink* pp. 199–217.

Stamatakis, A., 2014 RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stolarczyk, J. and J. Janick, 2011 Carrot: History and Iconography. *Chronica Horticulturae* **51**: 13–18.

Turner, S., 2014 *qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots*. biorXiv DOI: 10.1101/005165.

Tzuri, G. *et al.*, 2015 A 'golden' SNP in CmOr governs the fruit flesh color of melon (*Cucumis Melo*). *The Plant Journal* **82**: 267–279.

Vavilov, N. and V. Dorofeev, 1992 Origin and Geography of Cultivated Plants. (Cambridge [England] ; New York, NY, USA : Cambridge University Press), English edition.

Vos, P. *et al.*, 2017 Evaluation of ld decay and various ld-decay estimators in simulated and snp-array data of tetraploid potato. *Theoretical and Applied Genetics* **130**: 123–135.

Weir, B. and C. Cockerham, 1984 Estimating f-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

Wickham, H., 2009 *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York) .

Wright, S., 1951 The genetical structure of populations. *Ann.*

Eugenics **15**: 323–354.

Yan, J. *et al.*, 2009 Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLOS ONE* **4**: e8451.

Yuan, H., J. Zhang, D. Nageswaran, and L. Li, 2015 Carotenoid metabolism and regulation in horticultural crops. *Horticulture Research* **2**: hortres201536.

Zhao, K. *et al.*, 2011 Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nature Communications* **2**: 467.

Zhou, X. *et al.*, 2015 Arabidopsis OR proteins are the major post-transcriptional regulators of phytoene synthase in controlling carotenoid biosynthesis. *Proceedings of the National Academy of Sciences USA* **112**: 3558–3563.

Figure captions

Figure 1 Carrot accessions exhibiting the range of phenotypes used in this study and the stages of carrot domestication and improvement. From L to R: (A) Wild, (B) Eastern Landrace, (C) Western Historic Open Pollinated, (D) Modern Hybrids (L: Processing type; R: Imperator type). Photo courtesy of Matthew Mirkes.

Figure 2 Population structure of 463 carrot accessions with $< 30\%$ admixture (D2-LowAd). A) STRUCTURE groups. Percentage of membership (q) for each group identified at K=6. B) Geographic distribution of accessions each represented by a point on the map colored according to STRUCTURE group. Current commercial varieties not shown. C) PCA plot of the first two principal components. PC1 and PC2 account for 12.4% and 4.6% of the total variation, respectively. D) Maximum-likelihood tree of carrot accessions. Numbers on the branches indicate bootstrap support. Black branch represents outgroup *D. syrticus*. E) Color key. Total number of accessions in each STRUCTURE Group

Figure 3 Genome-wide association analysis of orange pigmentation and identification of the candidate gene *Or* on chromosome 3. A) Manhattan plot for orange carrot root color. Orange SNPs, with empirically-adjusted p-values less than 0.05, were defined as significant. B) Open reading frame of *Or* and the nonsynonymous mutation in exon 5 at position 3350 (T3350C). C) Box plots for lutein, alpha-carotene, and beta-carotene for the three *Or* genotypes (C/C, T/C, and TT) at position 3350. Center line = median, box limits = upper and lower quartiles, whiskers = 1.5× the interquartile range, dots = outliers. Different letters indicate significant differences between genotypes ($P < 0.05$, Tukey's HSD).

Figure 4 Regions of the carrot genome that likely underwent a selective sweep during domestication. A) Venn diagram represents the overlapping of 500 kb regions tested for selection signatures - top 5% of F_{st} and nucleotide diversity difference between wild and domesticated carrot accessions and top 1% of XP-CLR values. B) Genomic location of potential selective sweeps identified by F_{st}, nucleotide diversity and XP-CLR. The asterisk signifies the genome region carrying the candidate orange pigmentation gene, *Or*. C) Genome-wide linkage disequilibrium averaged across sliding windows of 100 SNPs in domesticated carrots. Regions identified as significant in A and B are highlighted in orange. The region containing the *Or* candidate gene for orange pigmentation in carrot is marked 'Or'.

Supplementary Table Captions

Table S1 Population summary of carrot accessions used in this study. Missing data is shown with a dot. Population type is either open pollinated (OP) or hybrid (H). Phloem color is White (w), Yellow (y), Red (r), Orange (o) or Purple (p). Accessions belonging to one of the STRUCTURE groups identified with Dataset D2-lowAd are color coded according to Fig. 2 in the main text.

Table S2 GWAS phenotype for pigment. 1 = orange; 0 = not orange; -999 = missing

Table S3 HPLC results for carotenoids and *Or* allele status in geographically distributed domesticated and wild carrot samples.

Table S4 Accessions used in *Or* alignment.

Table S5 Primer sequences used for amplifying the *Or* allele.

Table S6 Selective Sweep Statistics. The top 5% of values calculated for nucleotide diversity difference (> 1.58) and F_{st} (> 0.29) between wild and domesticated samples and top 1% for XP-CLR (> 11.94) in 500 kb sliding windows across the carrot genome are shown. Regions significant across all three tests are highlighted in orange. Previously described candidate domestication or improvement genes are listed in their corresponding genomic bin.

Table S7 Pairwise F_{st} between the six STRUCTURE groups. The lower triangle presents pairwise F_{st} values calculated with dataset D2. The upper triangle presents pairwise F_{st} values calculated from D2-lowAd. Increasing differentiation is shown with increasing red shading.

Table S8 Observed heterozygosity (H_o), within population gene diversity (H_s), overall gene diversity (H_t) and overall F_{st} averaged over all polymorphisms for datsets D1, D2 and D2-lowAd

Table S9 Annotated genes, via SwissProt, trEMBL, and Pfam, within the 143 kb GWAS signal. The carrot homolog of *Or* is highlighted in orange.

Supplementary Data

Sup. Data-D1 SNP file in variant call format (VCF) for Dataset D1.

Sup. Data-D2 SNP file in variant call format (VCF) for Dataset D2.

Supplementary Figures

Sup. Fig. S1

Average SNP density in 500 kb bins across the nine chromosomes for Dataset D1. Blue line is genome wide average of 54 SNPs/500 kb.

Sup. Fig. S2

Average SNP density in 500 kb bins across the nine chromosomes for dataset D2. Blue line is genome wide average of 43 SNPs/500 kb.

Chapter Three:

Comparison of representative and “custom” methods of generating core subsets of a carrot (*Daucus carota)* germplasm collection

Abstract:

Many breeding programs are interested in using genetic resources but have difficulty identifying accessions from germplasm collections because data that would be relevant to the program is missing or sparse. To efficiently use the diversity present in large germplasm collections, breeders often identify a subset of accessions that represents the larger collection. Methods for creating these “core collections” rely on partitioning collections into sub-clusters based on geographic, morphologic or genetic similarity. These methods do not consistently capture functional diversity and may be insufficient for breeder’s needs. Here, we use a collection of domesticated carrot (*Daucus carota*) accessions to compare representative methods with custom strategies that will allow breeders to create subsets of germplasm collections that maximize genetic diversity and trait values of interest. We find that for this collection, representative strategies are effective in capturing the diversity of the collection but do so no better than a random sample, likely because the collection itself is not strongly subdivided. Custom strategies that maximize genetic diversity and predicted trait values differ from the total collection with altered genetic, geographic and phenotypic compositions.

Introduction:

Plant breeders who want to increase the genetic diversity in their programs must make a challenging decision: with a genetically diverse group of crop accessions for which there is only imperfect and incomplete data, how to go about making strategic choices regarding which accessions to prioritize? Historically, researchers have worked to create “core collections”. These cores are meant to be representative, minimally redundant subsets of an entire collection. They can be screened for traits of interest and either used directly in a breeding program or can be used to direct researchers to other valuable entries in a collection. However, their development and use have been fraught with challenges.

Brown (1989) showed that it was theoretically possible to construct a core collection that maintained the allelic diversity of an entire collection. His mathematics require a collection to be genetically admixed, with the distribution of alleles uniform across the collection. In practice, collections rarely conform to such expectations. Strategies have been developed to first stratify collections into smaller subgroups with the expectation that these groups more closely resemble an ideal collection. From these subgroups, representative samples can be chosen.

Geographic origin, morphological descriptors, agronomic performance and neutral genetic markers have all been used – both alone and in combination – to construct core collections for many species. However, these strategies for developing cores are not sufficient to meet the needs of breeders. Often the variables used to stratify a collection are not predictive of diversity in other traits (Jansky et al 2015).

Furthermore, balancing overall diversity may fail to include important traits for breeding in the core.

Assessment of the strengths and weaknesses of various strategies to develop core collections has been limited by the lack of high-quality genomic data on entire collections. Here, we leverage the existing genetic resources available for carrot (*Daucus carota* L.) in order to explore the different strategies of developing core collections. Carrot is an outcrossing biennial species (2n=18) of nutritional importance, providing significant provitamin A. It is an attractive choice of model crop because it has significant genomic resources, including a sequenced genome. Carrot has received relatively little breeding attention with most efforts focused on increasing sweetness and beta-carotene content in elite processing lines. Therefore, strategies to incorporate genetically diverse material into carrot breeding programs will help advance breeding goals in other traits and market classes.

In this study, genetic, phenotypic and passport data on geographic origin are used to stratify a collection of 433 diverse Plant Introductions (PIs). Core collections are created by sampling from within stratified groups and then the representativeness of these cores is compared using various metrics.

We include two methods that do not first stratify the collection in our study. The first, hereafter referred to as Core Hunter core, is based on the Core Hunter algorithm designed by Thachuk et al (2009) which directly optimizes the genetic diversity of a core set. The second uses model-based prediction to identify accessions with high estimated trait values and is referred to as the genomic breeding values (GBV) core in this study.

For our purposes, we chose accessions for the GBV core with high predicted plant height and flavor scores, because these traits are important for carrot growers and consumers. Plant height is related to top vigor, which is important for weed control and mechanical harvest. Many elite lines do not have the vigorous tops desired by growers, so this is a trait that may benefit from incorporation of genetic resources with stronger tops. Good flavor improves the marketability; while modern elite carrot varieties are quite sweet and mild tasting, many historic lines suffer from harsh flavor notes. Identifying genetic resources with good flavor profiles will reduce the time needed to get back to an elite flavor profile.

While the neither the Core Hunter nor GBV core strategy is meant to be representative of the whole collection, they may be used directly to achieve specific breeding goals. They have the added benefit that they can useful without the initial requirement that the breeder extensively surveys a whole collection.

We show that for our collection of carrot PIs, strategies designed to choose representative core sets adequately represent the geographic, genetic and phenotypic diversity in the whole PI collection but a simple random sample does an equivalently good job. In contrast, our exploration of the GBV and Core Hunter cores reveals differences in composition that may recommend their use in breeding programs.

Methods:

Plant Material and Evaluation:

Four hundred thirty-three (433) cultivated *Daucus carota* PIs from the United States Department of Agriculture's National Plant Germplasm System (USDA-NPGS), maintained at the North Central Regional Plant Introduction Station (NCRPIS) in Ames, Iowa were included in this study. PIs were planted in a replicated trial (n.reps=2) at Hancock Research Station in Hancock, Wisconsin in the summer of 2016 and 2017. 250 seeds of each PI were planted in 1m rows. Plant height and width were measured twice during the season: at each time point three measurements were taken per plot. Disease severity was recorded late in the season on an ordinal scale (where 0=no disease, 5=100% diseased). Flavor, comprised of harshness and sweetness ratings on 0-5 scales, was evaluated on individual roots once by Dr. Phil Simon on a 0-5 scale (where 5=favorable flavor i.e. high sweetness or low harshness and 0=unfavorable flavor i.e low sweetness or high harshness). Stand count, the number of plants established per plot, was recorded early in the season.

Least-square phenotype means for each trait were estimated for each PI by fitting a linear mixed-effects model of the form:

$$y_{ijk} = \mu + g_i + year_j + rep(year)_{jk} + (genotype \times year)_{ij} + (genotype \times rep(year))_{ijk} + \varepsilon_{ijk}$$

where y_{ijk} is the trait measurement for PI *i*, year *j* and replicate (rep) *k*; μ is the grand mean; G_i is the fixed effect (genotypic value) of PI *i*; $year_j$, $rep(year)_{jk}$, $(genotype\times year)_{ij}$ and $(genotype\times rep(year))_{ijk}$ are the random effects of year j, rep *k* within year j, interaction between PI *i* and year *j*, and interaction between PI *i* and rep *k* within year j ; ε_{ijk} is the error. Random effects were modeled as independent and identically normally distributed. The model was fitted by restricted maximum likelihood (REML) using the R package lme4 (Bates et al., 2014).

To develop the GBV core a dataset of 145 commercially available carrot cultivars (CV) collected in 2013 (Luby et al 2016) and 273 open-pollinated (OP) cultivars collected before 1985 were used (Theisen 2016). Details regarding data collection for these two collections can be found in their respective publications.

Passport and phenotypic data for all accessions used in this study can be found in **Supplementary Table 1** (Appendix B).

Genotyping and SNP production

For the PIs and commercially available cultivars, total genomic DNA of individual plants was isolated from approximately 2g of lyophilized leaves of four-week old plants following the 10% CTAB protocol described by Murray and Thompson (Murray and Thompson. 1980) with modifications by Boiteux et al. (Boiteux et al.1999). The same protocol was applied to pooled samples of 8-12 plants of the OP cultivars. All DNA was quantified using the Quantus PicoGreen dsDNA Kit (Life Technologies, Grand Island NY) and normalized to 10ng/ul.

Genotyping-by-Sequencing (GBS), as described by Elshire et al. (Elshire et al. 2011), was carried out at the University of Wisconsin, Madison Biotechnology Center, (WI, USA) with minimal modification and half-sized reactions. Briefly, DNA samples were digested with ApeKI, barcoded and pooled for sequencing, and 80-95 pooled samples were run per single Illumina HiSeq 2000 lane, using paired end, 100 nt reads and v3 SBS reagents (Illumina, San Diego, CA). Images were analyzed using CASAVA 1.8.2. and bcl2fastq-1.8.4.

The TASSEL-GBS pipeline version 5.2.26 was used to call SNPs as described by Bradbury et al. (Bradbury et al. 2007) and Glaubitz et al. (Glaubitz et al. 2014) using the carrot reference genome (GenBank accession LNRQ01000000.1; Iorizzo 2016). Individual samples of the same PI were merged before SNPs were called. The SNP dataset had less than 10% missing data for genotype and marker, 10% minor allele frequency, and max minor allele frequency 0.05 leaving 19944 SNPs and 749 genotypes.

Methods of creating representative core collections:

Most methods of creating core collections begin by grouping like accessions and then taking samples from within those groups with the intention of developing a representative subset that avoids oversampling similar accessions. Three common strategies, stratifying by geographic origin, by genotypic distances, and by random sampling were evaluated in this study. Additionally, we explored a method that stratifies a collection by phenotypic distances but determined that it was not informative for our dataset so phenotypic stratification was not evaluated further. Sampling from within clusters to compose a core set introduces a degree of randomness, therefore 100 repetitions of each method were performed.

Random:

100 repeated random samples comprising 10% of the PI dataset (n=43) were generated in R.

Geographical origin:

Country of origin or collection site information for each PI was converted to approximate GPS coordinates using the web service HampsterMaps (n.d). A geographic distance matrix was generated using a Geographic Distance Matrix Generator published by the American Museum of Natural History, Center for Biodiversity Conservation (Ersts, n.d) which uses a set of spherical functions in order to calculate distance directly from geographic coordinates. Accession were clustered using the hclust function in the R packages stats (R core team). Ward's method, a hierarchical agglomerative clustering technique that minimize the total within-cluster variance, was found to produce comprehensible geographic clusters at K=6. 10% of each cluster was randomly sampled, and sampled PIs from each cluster were aggregated to form one geographic core. This was repeated 100 times.

Genetic diversity:

Genetic distances between PIs were calculated in TASSEL (Bradbury et al. 2007) using 19944 SNPs. Distance was defined as 1-IBS with IBS referring to the probability that alleles from a single locus drawn at random from two individuals are the same. Both Wards and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical clustering methods were used to cluster accessions according to their distance matrix. The correlation between an input distance matrix and a matrix of cophenetic distances (CPCC) — the distances at which two accessions are first joined in a dendrogram —provide evidence the presence or absence of genetic subgroups in a dataset (Odong et al. 2013). CPCC revealed little evidence of genetic subgrouping in the dataset; the carrot PI collection appears to more closely conform to the expectations set out by Brown (1989) than is typical of other collections (see Olivera et al. (2010), Skroch et al. (1998)) so stratification may not be necessary to construct a representative core. For the purposes of comparison, the *cuttree* function from the R stat package (R core team) was used to split PIs clustered using Ward's method into five subgroups. 10% of each cluster was randomly sampled, and sampled PIs from each cluster were aggregated to form one genotypic core. This was repeated 100 times.

Phenotypic diversity

Principal component analysis was performed on centered and scaled phenotypic values for 433 PIs using the FactoMineR package in R (Husson 2017). All PIs formed one large cloud when plotted on the first two principal components, which explained over half the variation in the dataset. Based on the results of this calculation, a phenotypic core was not created.

Methods of creating custom core collections:

Custom cores are not meant to be representative of the entire collection, rather they are optimized for some given criteria. The GBV core was designed to include PIs with high predicted values for plant height and flavor while the Core Hunter core maximizes the genetic distance between PIs in the set. Because they are based on optimization rather than sampling, it would be redundant to perform repetitions of the following two methods.

Genomic Breeding Values (GBV):

Top height and flavor phenotypes collected from the OP and CV collections were used in combination with an additive relationship matrix of all accessions to predict trait values for each accession. The additive relationship matrix was estimated as $A = \frac{ww'}{c}$ where $W_{ik} = X_{ik} + (1 - 2p_k)$ and p_k is the frequency of the allele at marker k and the EM imputation algorithm was used to estimate missing markers. Genomic-estimated breeding values for each PI were calculated for each trait using the kinship-based method in rrBLUP (Endelman 2011), which solves equations of the form $y = X\beta + [Z0]g + \epsilon$ with β as a vector of fixed effects, X is a design matrix for the fixed effects, g as a vector of random genotypic values which are $N{\sim}(0, K\sigma_u^2)$ when K is the additive relationship matrix, Z is a design matrix for the random effects and ε a vector of residuals which are normal with constant variance. Year, location, rep and reseeded were included as fixed effects. To achieve a set approximately equal to 10% of the total collection, 14 PIs were selected for each trait based on their GEBV and an equally-weighted index of the traits was calculated to select an additional 14 PIs. These accessions were assembled to form a balanced subset of 38 accessions (4 accessions were selected twice).

Optimization of genetic diversity (Core Hunter):

Core Hunter is a local search algorithm that generates representative subsets of a large dataset by optimizing different evaluating measures applied to a given distance matrix (Thachuk et al 2009). The function sampleCore() in the R version of Core Hunter (De Beukelaer 2017) was run on a precomputed genotypic distance matrix of 433 PIs.

Distance was calculated as 1-IBS with IBS referring to the probability that alleles from a single locus drawn at random from two individuals are the same. This function maximized the genotypic entry-to-nearest-entry distance for a 10% core subset of 43 PIs. Maximum time without improvement was 10 seconds by default.

Methods of comparing collections

We were interested in determining the representativeness of each core to the total collection and in parsing the differences between cores. Specifically, we were interested in how the custom cores differed from the representative cores. We examined the geographic, genotypic, and phenotypic composition of each core as well as the cores' ability to predict collection trait values when used as a training population in a genomic prediction model. For representative cores, comparison metrics were calculated on each repetition separately, unless otherwise noted, and a mean and standard deviations are reported on a per method basis.

Geographic representativeness

For each core, the count of PIs in each geographic cluster was calculated. Core counts (n=43, 38) from each cluster were compared to the number of individuals in each cluster of the entire collection (n=433) using Fisher's exact test in R, which is an appropriate test of independence of categorical data when some of the counts in each category are small. Using ggplot2 (Wickham et al 2016) in R, representative geographic maps were PIs were plotted according to their approximate geographic origin and geographic cluster identity.

Genotypic Representativeness

Population structure of each core and the entire collection were assessed using multidimensional scaling (principal coordinate analysis) of the genetic distance (1-IBS) matrix for N=433 individuals. K=2 dimensions were plotted for representative samples. Following the scoring method described by Noirot et al (1996), the contribution of each individual to the generalized sum of square of its set was calculated as the sum of squares of its K coordinates. The representativeness of a given subset was determined by the sum of the relative contributions of its members to the GSS of the whole set. Methods outlined by Odong et al., (2013) to assess measures of genetic distance among accessions in a core subset and between accessions in a core subset and those in the whole collection were calculated. These were the average distance between each accession in the full collection and the nearest entry in the core (ANE), average distance between each entry and the nearest neighbor entry (ENE) and average genetic distance between entries in the core (EE). Minor allelic frequencies, allelic richness, and observed and expected heterozygosity were calculated on a per locus basis using the R package hierfstat (Goudet 2014). T-tests for significant differences in overall measures of diversity were performed for each core sample. A Bonferroni correction for multiple tests were used to determine conservative significance levels. This research was performed in part using the computing resources and assistance of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences.

Phenotypic Representativeness

Using base R functions, phenotypic range, means, and variances were calculated for all traits in each custom core. For the other cores, the same statistics were calculated for each sample separately, and then aggregated. Trait correlation were calculated within the entire collection and within each core using Pearson's correlation coefficients with missing values deleted. Two-sided F-tests for significant differences in variances and T-tests for significant differences in means between the full and core collections were calculated.

Genomic Predictive Value

The value of using each core as a training population to predict trait values in the entire collection was evaluated using kin.blup() in the R package rrBLUP (Endleman 2011). Kinship for the predictive model was specified according to additive relationship matrix in which missing data was imputed using the EM algorithm. Accuracy was assessed as the correlation between least-square estimated phenotypes and predicted phenotypes for the full collection, minus those that were used in the training population.

Results and Discussion:

Development of representative cores

Geographic, phenotypic and genetic diversity was explored in a collection of 433 PIs. Following the general method outlined by Frankel and Brown, the collection was clustered into like groups which were used to guide the development of representative cores.

Geographic diversity

Hierarchical cluster analysis performed on geographic origin data revealed the presence of six well-distributed geographic clusters with accessions grouping into Central Asian (57), Northern European (153), South-Eastern Europe/Middle East/North Africa (100), Eastern Asian (71), United States (49) and New Zealand (3) clusters **(Figure 1, Table 3).**

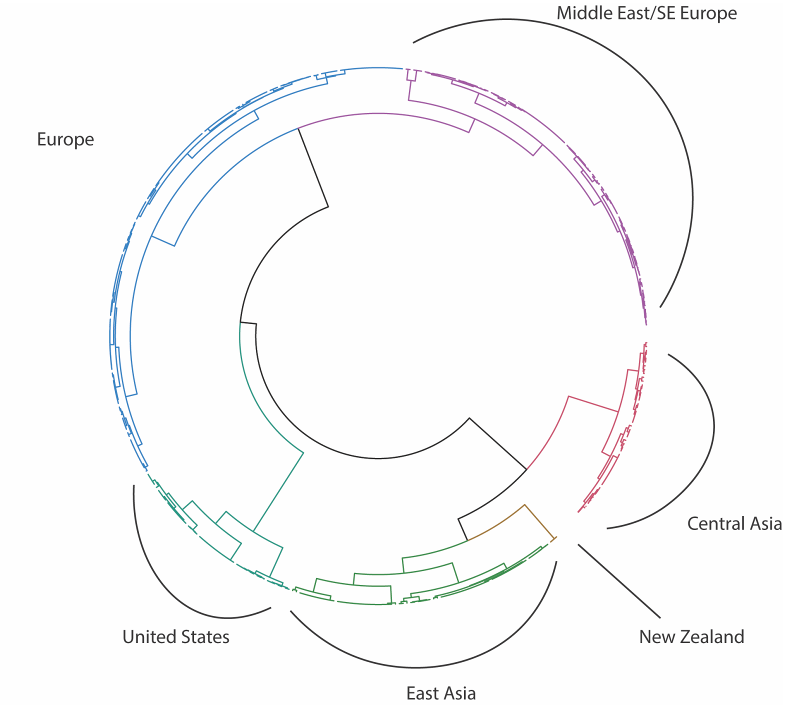
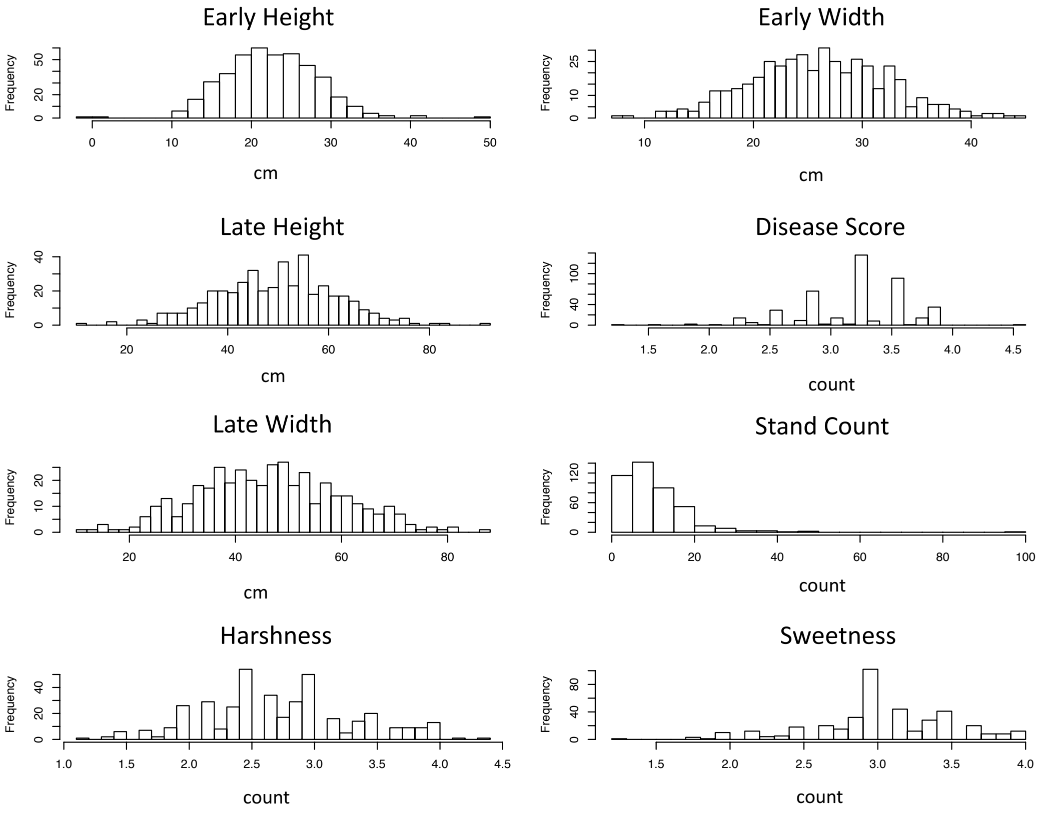


Figure 1: Geographic clustering of 433 PIs. Ward's minimum variance criteria was applied to a geographic distance matrix calculated from estimated latitude and longitude data for 433 PIs. K=6 clusters grouped PIs according to geographic region.

Phenotypic diversity

Least square mean estimates for plant height and width were both approximately normally distributed **(Figure 2).** Scored traits had slightly more uniform distributions. Mean trait values are reported in **Table 1.** Pearson's correlation coefficient, which measures the linear correlation between quantitative variables, was moderately high for height and width throughout the season. Early season stand count was moderately correlated with early plant height but not height or width measurements later in the season. Disease score and flavor scores were weakly correlated with other traits **(Table 2).**

Principal component analysis on centered and scaled phenotypic values was performed **(Figure 3).** 44.83% of the variance in the dataset was explained by the first principal component, which was heavily controlled by height and width measurements. Flavor contributed to the second component, which explained 15.69% of the variation in the dataset. PIs plotted according to their coordinates for the top two components showed that hierarchical cluster analysis failed to reveal any interpretable clusters. It was determined that creating a core collection based on the phenotypes available for this collection would not be meaningful.



Early Height

Early Width

Late Height

Disease Score

Late Width

Stand Count

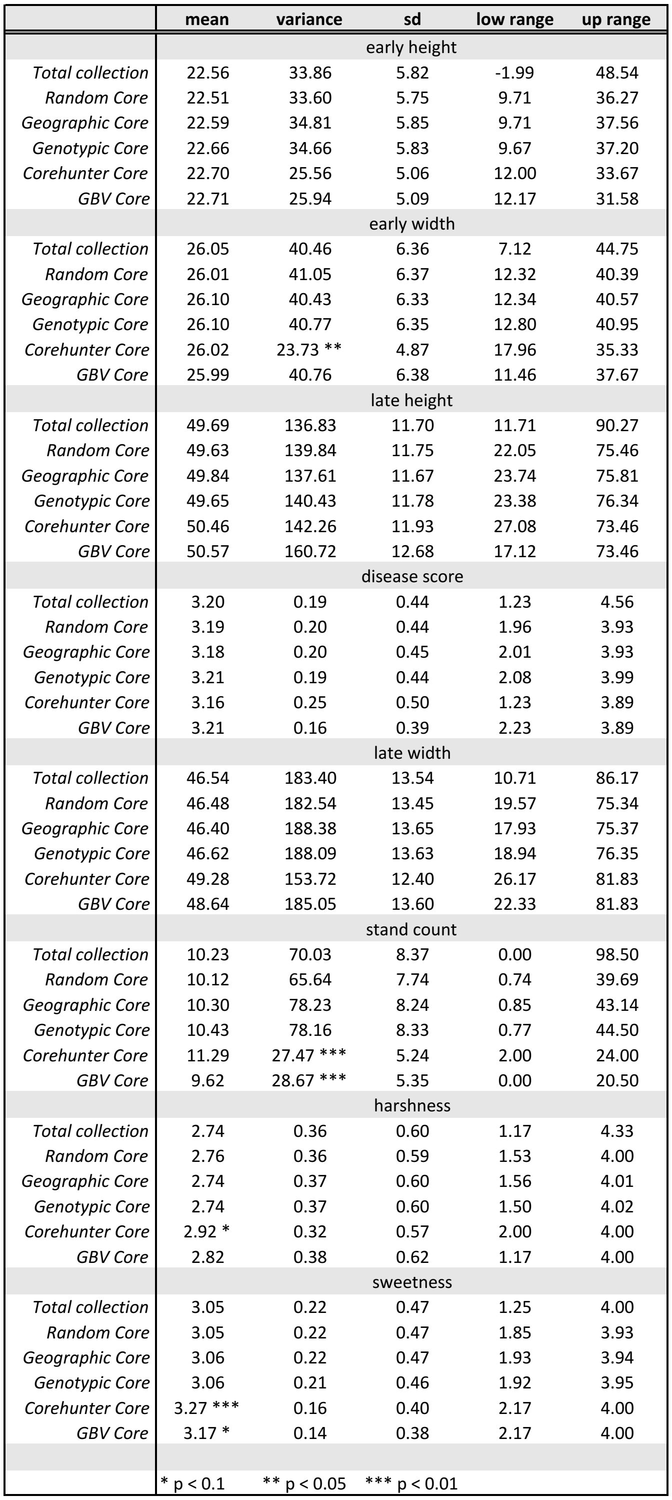
Harshness

Sweetness

Figure 2: Distribution of phenotypes for 433 PIs. 433 PIs were grown over two years in Hancock, Wisconsin with two replications. Least-square mean estimates of trait values, with within plot measurements averaged for height and width, are plotted.

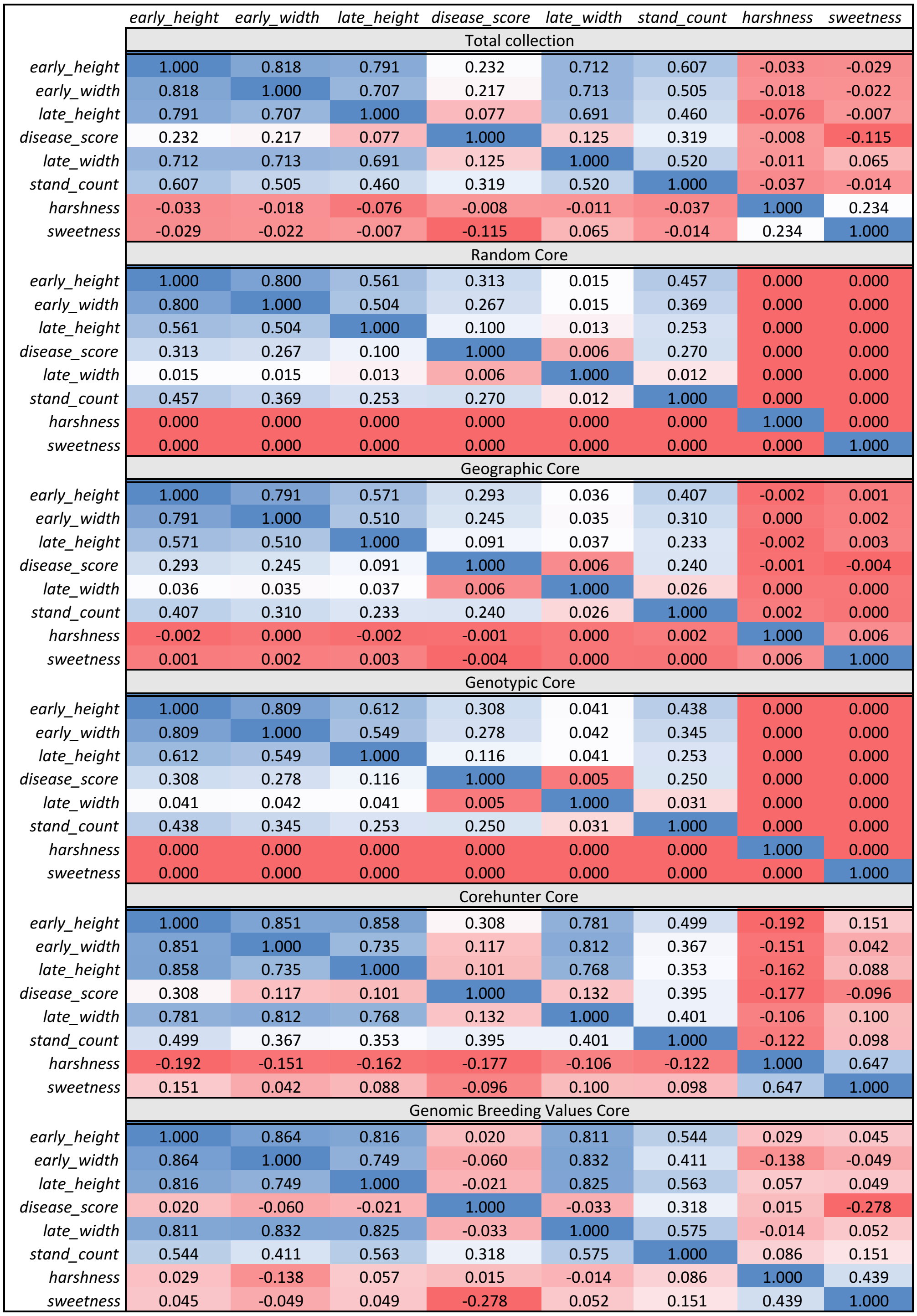
Genotypic diversity

Genetic diversity was measured on a reduced SNP dataset containing genotypes for 433 PIs. Overall observed heterozygosity was 0.302 and overall expected heterozygosity was 0.375 which indicates a moderate reduction in genetic diversity in this population compared to expectations at equilibrium. Overall minor allele frequency was 0.274 and allele richness was 2.01 **(Table 5).**



	mean	variance	sd	low range	up range
			early height		
Total collection	22.56	33.86	5.82	-1.99	48.54
Random Core	22.51	33.60	5.75	9.71	36.27
Geographic Core	22.59	34.81	5.85	9.71	37.56
Genotypic Core	22.66	34.66	5.83	9.67	37.20
Corehunter Core	22.70	25.56	5.06	12.00	33.67
GBV Core	22.71	25.94	5.09	12.17	31.58
			early width		
Total collection	26.05	40.46	6.36	7.12	44.75
Random Core	26.01	41.05	6.37	12.32	40.39
Geographic Core	26.10	40.43	6.33	12.34	40.57
Genotypic Core	26.10	40.77	6.35	12.80	40.95
Corehunter Core	26.02	23.73 **	4.87	17.96	35.33
GBV Core	25.99	40.76	6.38	11.46	37.67
			late height		
Total collection	49.69	136.83	11.70	11.71	90.27
Random Core	49.63	139.84	11.75	22.05	75.46
Geographic Core	49.84	137.61	11.67	23.74	75.81
Genotypic Core	49.65	140.43	11.78	23.38	76.34
Corehunter Core	50.46	142.26	11.93	27.08	73.46
GBV Core	50.57	160.72	12.68	17.12	73.46
			disease score		
Total collection	3.20	0.19	0.44	1.23	4.56
Random Core	3.19	0.20	0.44	1.96	3.93
Geographic Core	3.18	0.20	0.45	2.01	3.93
Genotypic Core	3.21	0.19	0.44	2.08	3.99
Corehunter Core	3.16	0.25	0.50	1.23	3.89
GBV Core	3.21	0.16	0.39	2.23	3.89
			late width		
Total collection	46.54	183.40	13.54	10.71	86.17
Random Core	46.48	182.54	13.45	19.57	75.34
Geographic Core	46.40	188.38	13.65	17.93	75.37
Genotypic Core	46.62	188.09	13.63	18.94	76.35
Corehunter Core	49.28	153.72	12.40	26.17	81.83
GBV Core	48.64	185.05	13.60	22.33	81.83
			stand count		
Total collection	10.23	70.03	8.37	0.00	98.50
Random Core	10.12	65.64	7.74	0.74	39.69
Geographic Core	10.30	78.23	8.24	0.85	43.14
Genotypic Core	10.43	78.16	8.33	0.77	44.50
Corehunter Core	11.29	27.47 ***	5.24	2.00	24.00
GBV Core	9.62	28.67 ***	5.35	0.00	20.50
			harshness		
Total collection	2.74	0.36	0.60	1.17	4.33
Random Core	2.76	0.36	0.59	1.53	4.00
Geographic Core	2.74	0.37	0.60	1.56	4.01
Genotypic Core	2.74	0.37	0.60	1.50	4.02
Corehunter Core	2.92 *	0.32	0.57	2.00	4.00
GBV Core	2.82	0.38	0.62	1.17	4.00
			sweetness		
Total collection	3.05	0.22	0.47	1.25	4.00
Random Core	3.05	0.22	0.47	1.85	3.93
Geographic Core	3.06	0.22	0.47	1.93	3.94
Genotypic Core	3.06	0.21	0.46	1.92	3.95
Corehunter Core	3.27 ***	0.16	0.40	2.17	4.00
GBV Core	3.17 *	0.14	0.38	2.17	4.00
	* p < 0.1	** p < 0.05	*** p < 0.01		

Table 1: Least-square estimated phenotypes for core sets and total collection. Significant differences in means and variances between core sets and total collection are indicated. For representative cores, values are averaged over 100 samples.



	early_height	early_width	late_height	disease_score	late_width	stand_count	harshness	sweetness
	Total collection							
early_height	1.000	0.818	0.791	0.232	0.712	0.607	-0.033	-0.029
early_width	0.818	1.000	0.707	0.217	0.713	0.505	-0.018	-0.022
late_height	0.791	0.707	1.000	0.077	0.691	0.460	-0.076	-0.007
disease_score	0.232	0.217	0.077	1.000	0.125	0.319	-0.008	-0.115
late_width	0.712	0.713	0.691	0.125	1.000	0.520	-0.011	0.065
stand_count	0.607	0.505	0.460	0.319	0.520	1.000	-0.037	-0.014
harshness	-0.033	-0.018	-0.076	-0.008	-0.011	-0.037	1.000	0.234
sweetness	-0.029	-0.022	-0.007	-0.115	0.065	-0.014	0.234	1.000
	Random Core							
early_height	1.000	0.800	0.561	0.313	0.015	0.457	0.000	0.000
early_width	0.800	1.000	0.504	0.267	0.015	0.369	0.000	0.000
late_height	0.561	0.504	1.000	0.100	0.013	0.253	0.000	0.000
disease_score	0.313	0.267	0.100	1.000	0.006	0.270	0.000	0.000
late_width	0.015	0.015	0.013	0.006	1.000	0.012	0.000	0.000
stand_count	0.457	0.369	0.253	0.270	0.012	1.000	0.000	0.000
harshness	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
sweetness	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
	Geographic Core							
early_height	1.000	0.791	0.571	0.293	0.036	0.407	-0.002	0.001
early_width	0.791	1.000	0.510	0.245	0.035	0.310	0.000	0.002
late_height	0.571	0.510	1.000	0.091	0.037	0.233	-0.002	0.003
disease_score	0.293	0.245	0.091	1.000	0.006	0.240	-0.001	-0.004
late_width	0.036	0.035	0.037	0.006	1.000	0.026	0.000	0.000
stand_count	0.407	0.310	0.233	0.240	0.026	1.000	0.002	0.000
harshness	-0.002	0.000	-0.002	-0.001	0.000	0.002	1.000	0.006
sweetness	0.001	0.002	0.003	-0.004	0.000	0.000	0.006	1.000
	Genotypic Core							
early_height	1.000	0.809	0.612	0.308	0.041	0.438	0.000	0.000
early_width	0.809	1.000	0.549	0.278	0.042	0.345	0.000	0.000
late_height	0.612	0.549	1.000	0.116	0.041	0.253	0.000	0.000
disease_score	0.308	0.278	0.116	1.000	0.005	0.250	0.000	0.000
late_width	0.041	0.042	0.041	0.005	1.000	0.031	0.000	0.000
stand_count	0.438	0.345	0.253	0.250	0.031	1.000	0.000	0.000
harshness	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
sweetness	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
	Corehunter Core							
early_height	1.000	0.851	0.858	0.308	0.781	0.499	-0.192	0.151
early_width	0.851	1.000	0.735	0.117	0.812	0.367	-0.151	0.042
late_height	0.858	0.735	1.000	0.101	0.768	0.353	-0.162	0.088
disease_score	0.308	0.117	0.101	1.000	0.132	0.395	-0.177	-0.096
late_width	0.781	0.812	0.768	0.132	1.000	0.401	-0.106	0.100
stand_count	0.499	0.367	0.353	0.395	0.401	1.000	-0.122	0.098
harshness	-0.192	-0.151	-0.162	-0.177	-0.106	-0.122	1.000	0.647
sweetness	0.151	0.042	0.088	-0.096	0.100	0.098	0.647	1.000
	Genomic Breeding Values Core							
early_height	1.000	0.864	0.816	0.020	0.811	0.544	0.029	0.045
early_width	0.864	1.000	0.749	-0.060	0.832	0.411	-0.138	-0.049
late_height	0.816	0.749	1.000	-0.021	0.825	0.563	0.057	0.049
disease_score	0.020	-0.060	-0.021	1.000	-0.033	0.318	0.015	-0.278
late_width	0.811	0.832	0.825	-0.033	1.000	0.575	-0.014	0.052
stand_count	0.544	0.411	0.563	0.318	0.575	1.000	0.086	0.151
harshness	0.029	-0.138	0.057	0.015	-0.014	0.086	1.000	0.439
sweetness	0.045	-0.049	0.049	-0.278	0.052	0.151	0.439	1.000

Table 2: Pearson's correlation coefficients for each trait in core sets and total collection. Incomplete observations are removed. Blue indicates high correlation between traits, red indicates low correlation between traits.

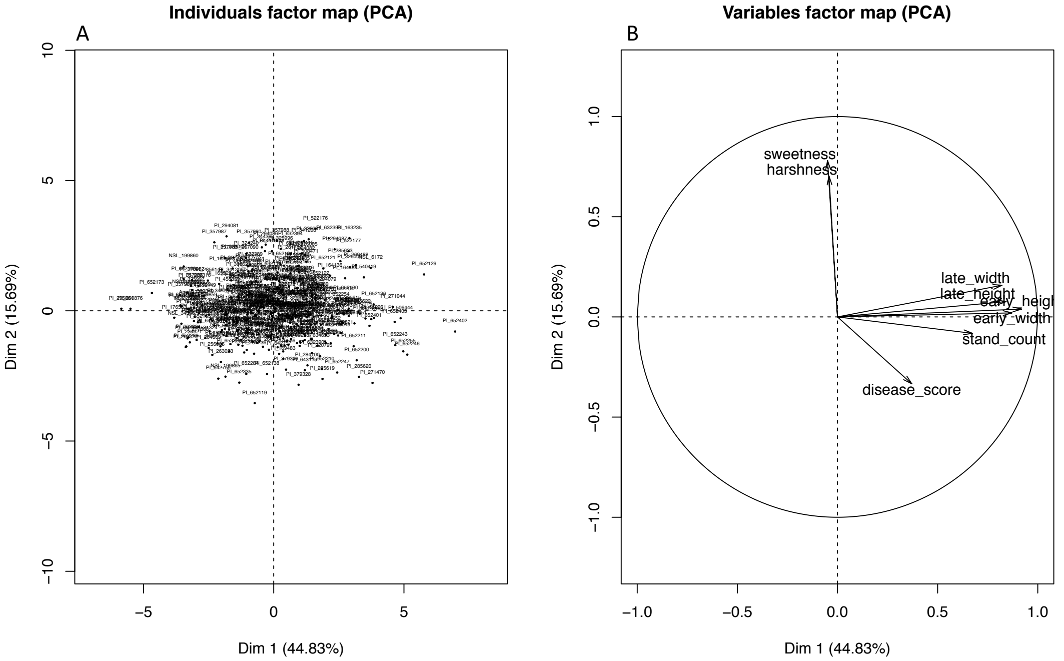


Figure 3: PCA of collection phenotype data: PCA was performed on scaled and centered LS-mean estimates (n=433) for collection trait values. A) Individual factor map plotting PI coordinates on first two principal components B) Variables factor map showing the projection of original variables onto the first two principal components

Principal coordinate analysis on a genetic distance matrix of the same 433 PIs, however, did not reveal obvious genetic subgroups **(Figure 7).** Hierarchical cluster analysis was also used to explore genetic subgrouping of the dataset. Two methods of clustering the dataset were compared, Ward's minimum variance method and UPGMA using the correlation between cophenetic distances (CPCC). Higher CPCC values provide stronger evidence the presence of genetic subgroups in a dataset (Odong et al. 2013). The CPCC values were 0.58 and 0.89 for dendrograms created with Ward's and UPGMA methods, respectively, which provides only weak evidence for genetic subgrouping **(Figure 4).** K clusters, where K was increased from 1-9, were plotted on a world map where points indicated geographic origin for a single PI and color indicated K cluster identity **(Figure 5).** Averages silhouette width was maximized at K=3 clusters (data not shown). In previous analysis, cultivated carrots have been found to cluster into Eastern and Western genetic subgroups, results which are recapitulated here when K=3. In the results presented here, K was set to 5, however there are no meaningful differences in core composition when K=5 vs. K=3 (data not shown).

<u>Development of custom cores</u>

Core Hunter

The Core Hunter algorithm (Thachuk et al 2009) was used to optimize the between entry genetic distance in a core set of 43 accessions. The maximum between entry genetic distance was 0.403.

Genomic Breeding Values

Using previously collected phenotype data (Luby et al 2016) (Theisien 2016), top height and harshness were predicted in the PI collection. Correlations between predicted and estimated trait values in the PI collection were 0.31 for top height and 0.11 for harshness. Accessions with top predicted trait values were selected for the GBV core set. While predictive ability of the model used to choose the GBV core was low, likely because previous phenotypes on a different set of cultivars were used to train the prediction equation, this strategy responds to realistic limitations breeders may face with regards to available phenotypes.

Comparison of core collections

Except in the case of the GBV collection, each core was chosen to represent 10% (n=43) of the full PI collection (n=433). The GBV collection was composed of 38 PIs (8.7%) due to redundancy in selections for certain traits. The representativeness of each core to the total collection was compared in terms of its genetic, geographic and phenotypic diversity and salient differences between each core were interpreted.

Geographic representativeness

Using the six geographic groups identified in hierarchical cluster analysis of the full PI collection, the geographic representativeness of each of the each of the core collections was analyzed (**Figure 6**). In four of the cores, the geographic representativeness was proportionate to the geographic distribution of the whole collection as determined via Fisher's exact test **(Table 3)**. The geographic distribution of the PIs in the GBV core differed significantly from the whole collection (p=0.024). This core underrepresented accessions from Central Asia and Southern Europe/MENA. This could be because the training population used to build the predictive mode wester overrepresented cultivated accessions or because mild, sweet flavor has been more strongly prioritized in western accessions.

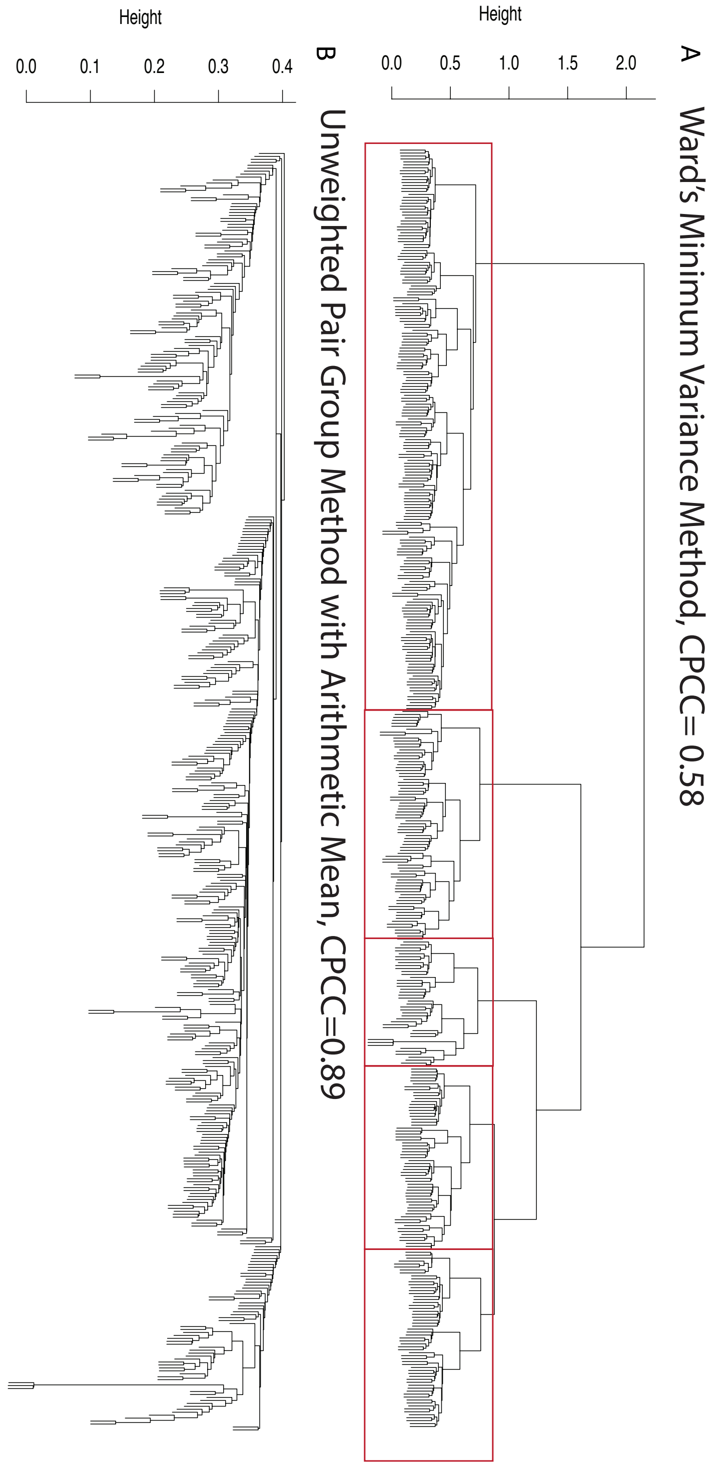


Figure 4: Agglomerative clustering of PI genotypes: Two methods are hierarchical agglomerative clustering performed on a genetic distance matrix (n=433) are compared A) Ward's minimum variance method and B) UPGMA. Co-phenetic correlation coefficients are reported and do not provide strong evidence for genetic subclusters. Red boxes in (A) indicate groups used in downstream analysis

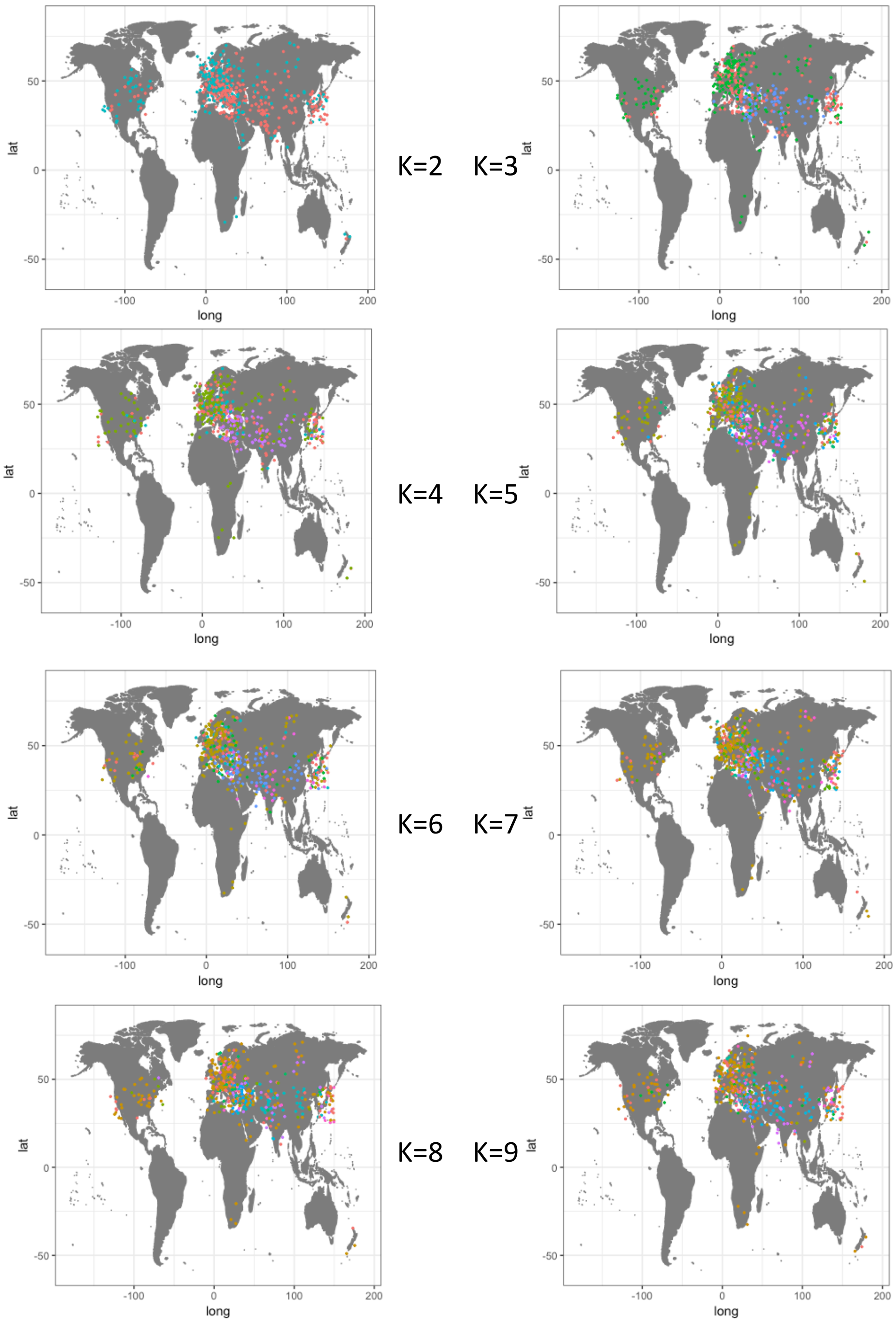


Figure 5: Genetic structure is weakly correlated with geographic origin. Ward's minimum variance criteria was applied to a genetic distance matrix (n=433) to construct a dendrogram that was then cut to construct K groups, where K was varied from 2 to 9. PIs are plotted according to their geographic origin and colored according to K group identity.

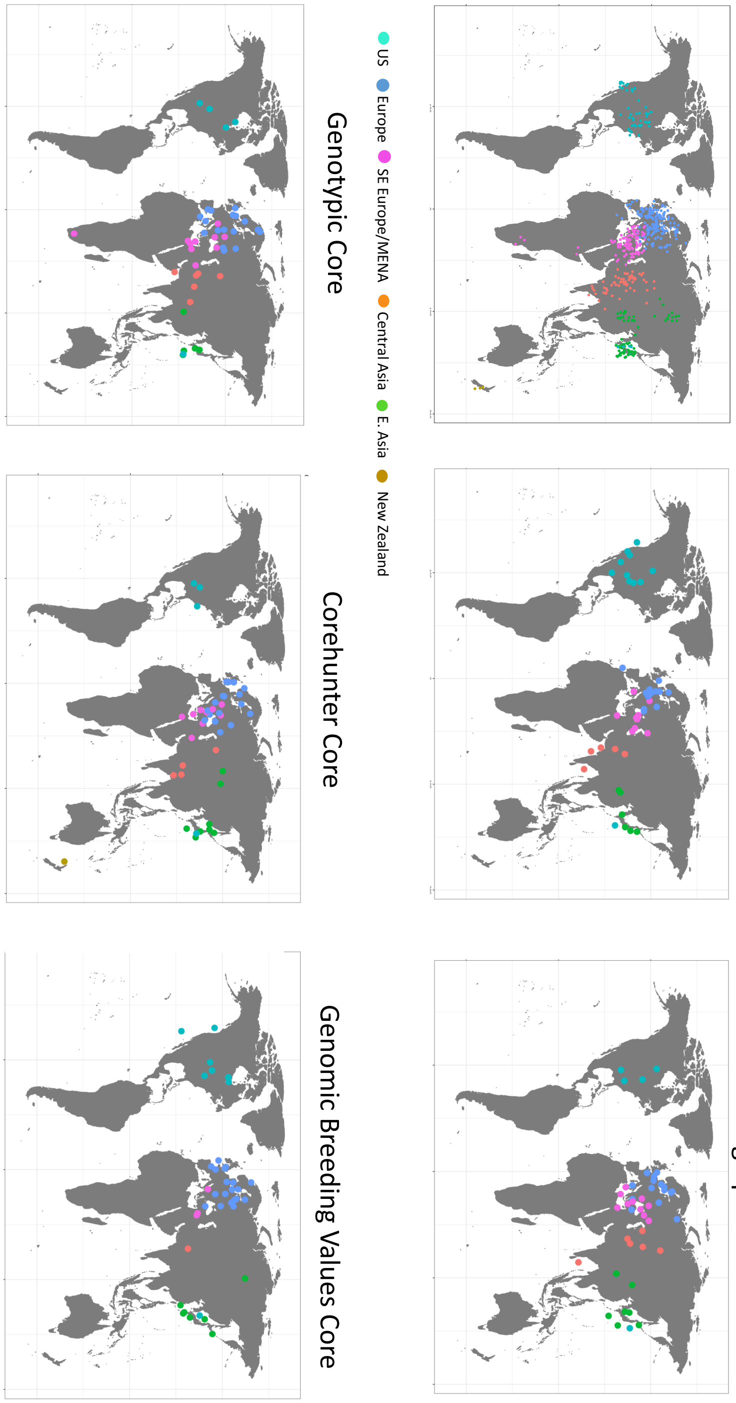


Figure 6: Maps of geographic origin for PIs in total collection and in representative cores: PIs are plotted on a world map according to their approximate geographic origin. Colors represent geographic subcluster identity

	Central Asia	N. Europe	SE Europe/MENA	E. Asia	United States	New Zealand	total N	Fisher's P-value
Total collection	57	153	100	71	49	3	433	NA
Random Core	6	16	10	7	5	1	43	0.85
Geographic Core	6	15	10	7	5	0	43	1.00
Genotypic Core	6	15	10	7	5	1	43	0.84
Corehunter Core	4	17	9	8	4	1	43	0.73
GBV Core	1	19	3	7	8	0	38	0.02

Table 3: Geographic representativeness of core sets. Number of PIs in each geographic subgroup, as defined by cluster analysis, per core (averaged over 100 samples for representative cores) and total collection. Differences between each core and total collection are compared via Fisher's exact test.

Phenotypic relatedness

The phenotypic means, variances and ranges were calculated for all cores **(Table 1)**. T-tests for differences in means and F-tests for differences in variances between cores and the full collection were calculated. The representative cores did not differ significantly from the whole collection for any of the traits measured.

Mean sweetness scores for the GBV core were significantly higher than the whole collection (p=0.076) and stand count (p=0.001) variances were lower than the full collection. The Core Hunter core had significant higher means for harshness (p=.058) and sweetness (p=0.001) (higher scores are desirable for both traits). Trait variances for the Core Hunter core differed significantly from the full collection for early plant width p=0.016) and stand count (p<.0001).

Given that the GBV core was composed of accessions selected according to their high GEBV for flavor and top height, it is rather surprising that this core did not have more extreme trait value related to the entire collection. High variances for these traits and relatively low predictive ability of the kinship model could result in the selection of individuals with moderate estimated phenotypes.

Representative core collections should also preserve correlation among traits. The magnitude of correlations between traits within a core follows a similar pattern to correlations within the whole collection for most traits and cores. ***(*Table 2)**. However, breeding programs often seek to change these correlations so an optimized collection may shift trait correlations. Ideally, a carrot variety would have both low disease scores and large plant height; these traits are moderately correlated in the total collection (0.07-0.232). In the GBV core, however, the correlation between disease score and plant height/width is decreased (-0.06-0.02) which could be advantageous in a breeding program.

Genotypic representativeness

Multidimensional scaling (K=2) was performed on a distance matrix of the whole collection of 433 PIs. The generalized sum of squares (GSS) of the whole dataset was 6.47. The sum of squares of the individuals in each core set was calculated and the principal component score (Noiroit et al 1996) for each core was found by dividing the core sum of squares by total GSS. A perfectly representative core subset composing 10% of the collection should have a PC score of 0.1. Principal component scores for the cores ranged from 0.10 for the geographic core to 0.049 for the Core Hunter core. The genotypic core and random core also represented the genetic diversity in the whole collection well, with PC scores near 0.1.

A plot of accessions in each core according to their PC coordinates shows that, compared to other methods of generating a core collection, the Core Hunter strategy sampled more accessions with moderately divergent genotypes, maximizing the overall distance between accessions but resulting in a lower PC score. The GBV core, with a PC score of 0.066, represents a midpoint between the Core Hunter strategy and the representative strategies which nonetheless appear to sample more extreme genotypes (**Figure 7**).

The degree to which a given core represents the diversity in a total collection can be further summarized by accounting for the genetic distances between accessions in the core and the whole collection **(Table 4).** Average distance between each PI in the full collection and the nearest entry in the core (ANE) ranged from 0.314 for the genotypic, geographic and random cores to 0.353 for the Core Hunter core. Average distance between each entry and the nearest neighbor entry (ENE) in the core ranged from 0.287 for the genotypic core to 0.387 for the Core Hunter core. Average genetic distance between entries in the core (EE) ranged from 0.375 for the geographic core to 0.40 for the core hunter core. Based on these measurements, the representative cores better represented the whole collection while the Core Hunter core maximized the diversity of the core itself. The GBV core again represents a midpoint between these two goals.

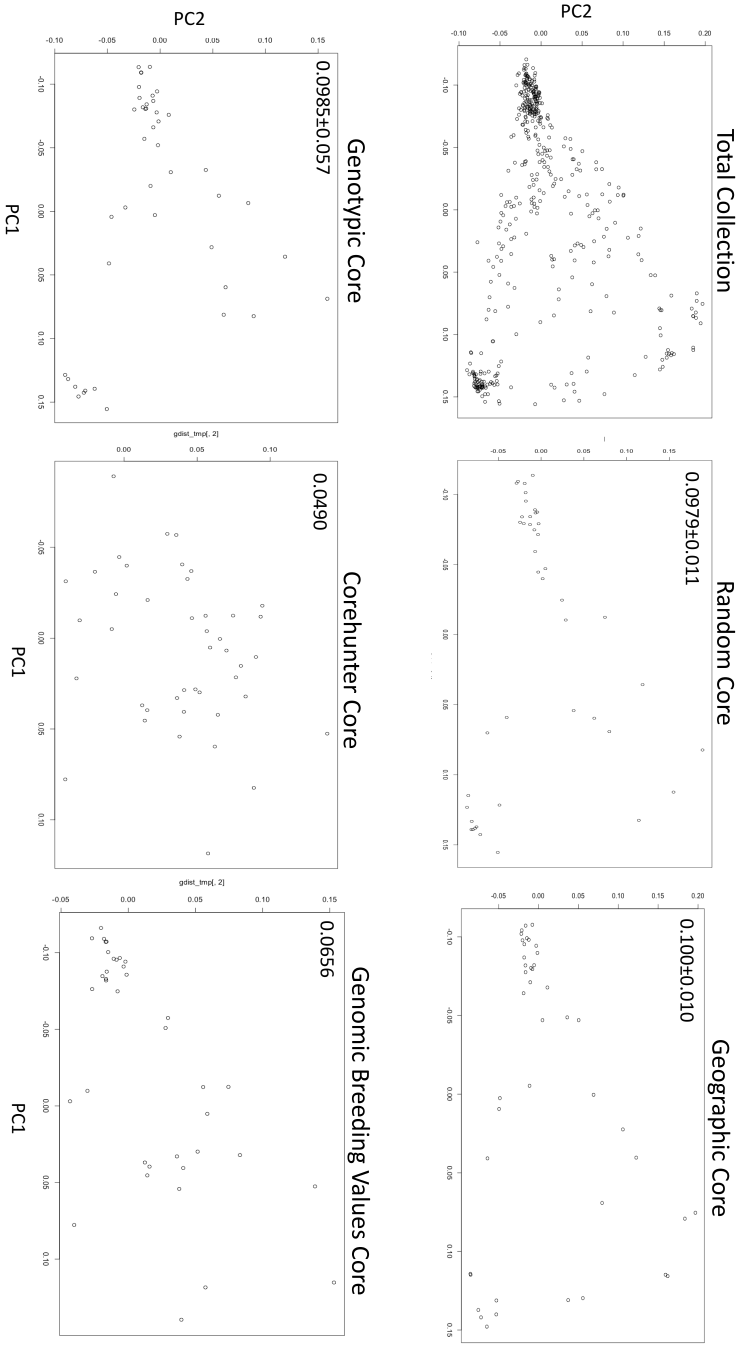
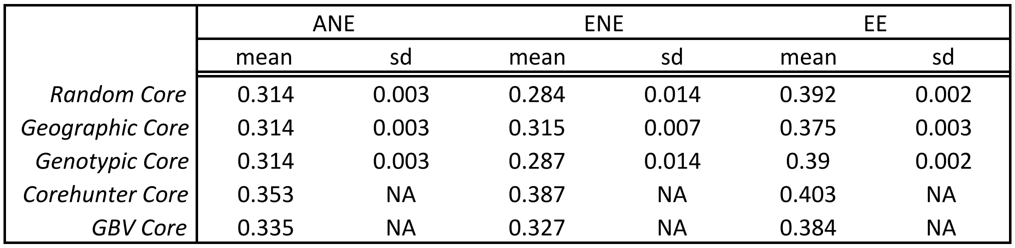


Figure 7: MDS plots of PIs in total collection and representative core. PIs are plotted according to their MDS coordinates. Mean and standard deviation of PC scores for each core are reported in upper left corner of each plot. PC score describes contribution of entries in each core to the generalized sum of squares (GSS) of the whole collection.

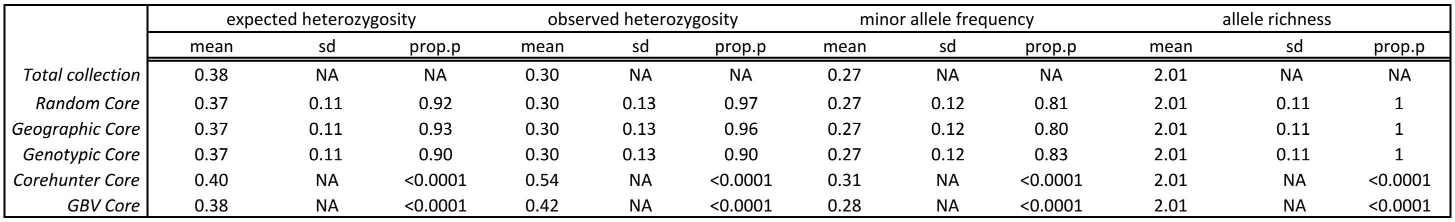


	ANE		ENE		EE	
	mean	sd	mean	sd	mean	sd
Random Core	0.314	0.003	0.284	0.014	0.392	0.002
Geographic Core	0.314	0.003	0.315	0.007	0.375	0.003
Genotypic Core	0.314	0.003	0.287	0.014	0.39	0.002
Corehunter Core	0.353	NA	0.387	NA	0.403	NA
GBV Core	0.335	NA	0.327	NA	0.384	NA

Table 4: Genetic representativeness of core sets. ANE measures the average distance between each PI in the full collection and the nearest entry in the core. ENE measures the distance between ach entry and the nearest neighbor entry within a core. EE measures the average genetic distance between entries in the core. Distances are defined as 1-IBS where IBS is identity-by-state.

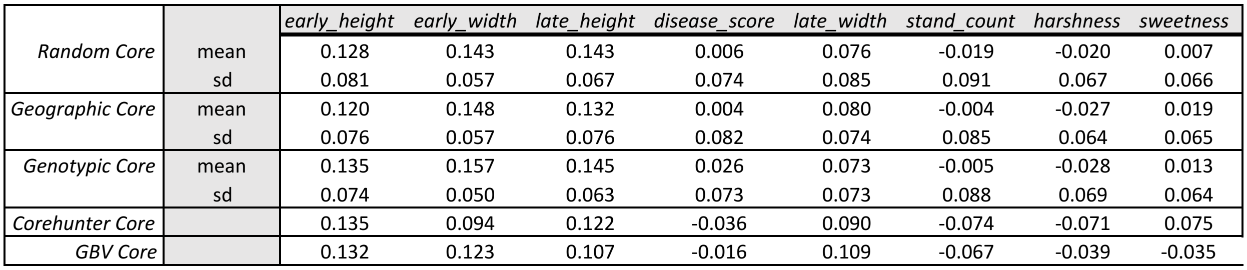
Measures of genetic diversity were also calculated for each core **(Table 5).** Expected heterozygosity ranged from 0.4 for the Core Hunter core to 0.37 for the representative cores. Observed heterozygosity was also highest in the Core Hunter core (0.54) and lowest in the same set of cores (0.30). Expected and observed heterozygosity were 0.38 and 0.42 for the GBV core. Like in the whole collection, expected heterozygosity was higher than observed heterozygosity for the three similar cores, but in the GBV and Core Hunter cores, observed heterozygosity was higher than expected. Minor allele frequency was 0.27 for the three similar cores and 0.31 and 0.28 for the Core Hunter and GBV cores. Allele richness was 2.0 in all cores.

All three of the representative methods represented the genetic diversity in the whole collection and did not seem to sacrifice rare alleles in the reduced subset. Conversely, the Core Hunter and GBV cores had altered patterns of genetic diversity compared to the whole collection (**Figure 8**). Compared to the other cores, they had higher minor allele frequencies and observed heterozygosity; these strategies increased the frequency of minor alleles in the core relative to the total collection.



	expected heterozygosity			observed heterozygosity			minor allele frequency			allele richness		
	mean	sd	prop.p	mean	sd	prop.p	mean	sd	prop.p	mean	sd	prop.p
Total collection	0.38	NA	NA	0.30	NA	NA	0.27	NA	NA	2.01	NA	NA
Random Core	0.37	0.11	0.92	0.30	0.13	0.97	0.27	0.12	0.81	2.01	0.11	1
Geographic Core	0.37	0.11	0.93	0.30	0.13	0.96	0.27	0.12	0.80	2.01	0.11	1
Genotypic Core	0.37	0.11	0.90	0.30	0.13	0.90	0.27	0.12	0.83	2.01	0.11	1
Corehunter Core	0.40	NA	<0.0001	0.54	NA	<0.0001	0.31	NA	<0.0001	2.01	NA	<0.0001
GBV Core	0.38	NA	<0.0001	0.42	NA	<0.0001	0.28	NA	<0.0001	2.01	NA	<0.0001

Table 5: Overall genetic diversity measures for core sets and total collection. Overall (mean) observed heterozygosity, expected heterozygosity, minor allele frequency and allele richness are recorded. For all measures on all cores, a test of significant differences between average per locus values and total collection per-locus values was significant at alpha=0.05. P-values on a per-core, per-locus basis were calculated for the same measures and the proportion of p-values less than 0.0005 (alpha corrected for multiple tests) is shown in the prop.p column. For custom cores, a single p-value is reported



		early_height	*early_width*	*late_height*	*disease_score*	*late_width*	*stand_count*	*harshness*	*sweetness*
Random Core	mean	0.128	0.143	0.143	0.006	0.076	-0.019	-0.020	0.007
	sd	0.081	0.057	0.067	0.074	0.085	0.091	0.067	0.066
Geographic Core	mean	0.120	0.148	0.132	0.004	0.080	-0.004	-0.027	0.019
	sd	0.076	0.057	0.076	0.082	0.074	0.085	0.064	0.065
Genotypic Core	mean	0.135	0.157	0.145	0.026	0.073	-0.005	-0.028	0.013
	sd	0.074	0.050	0.063	0.073	0.073	0.088	0.069	0.064
Corehunter Core		0.135	0.094	0.122	-0.036	0.090	-0.074	-0.071	0.075
GBV Core		0.132	0.123	0.107	-0.016	0.109	-0.067	-0.039	-0.035

Table 6: Genomic predictive ability of each core used as a training population for the total collection. Predictive ability is defined as the correlation between predicted and estimated trait values for the total collection minus those used in the training population.

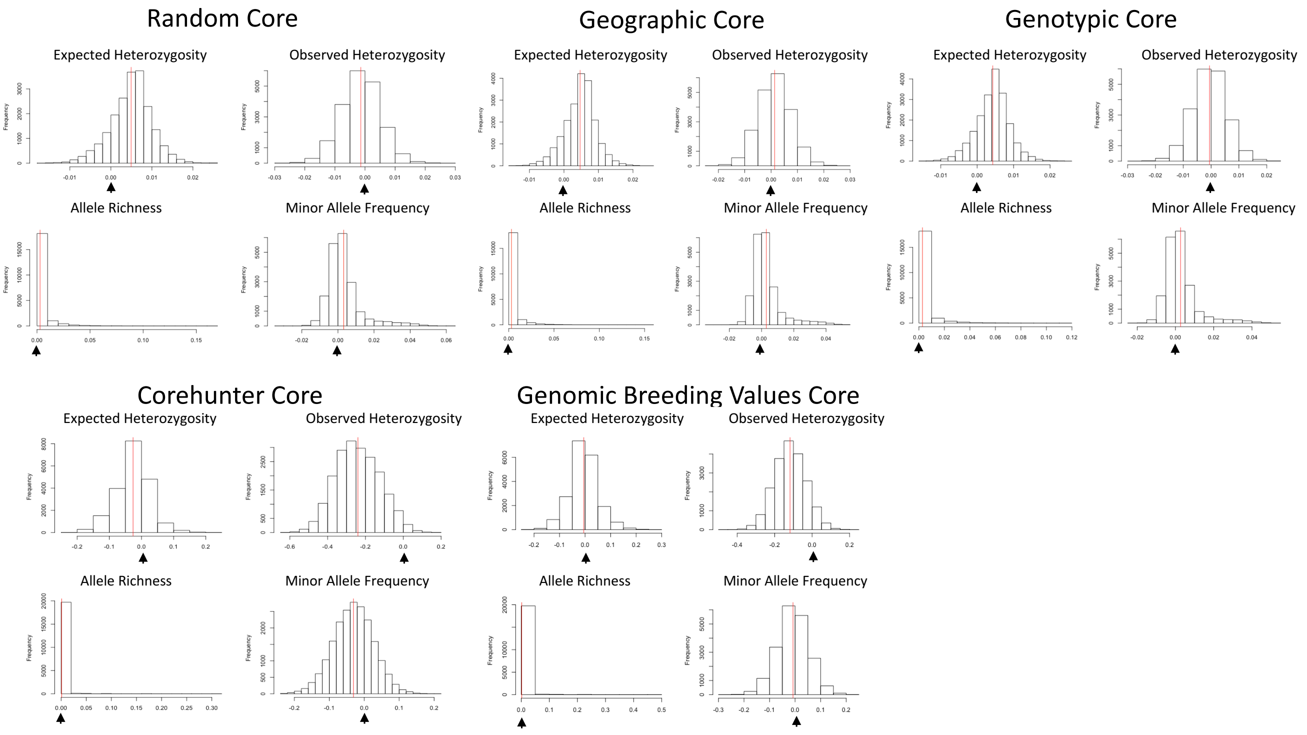


Figure 8: Per-locus distribution of differences in genetic diversity measurements between total collection and core. Row A) Differences are between average of 100 sampled cores and total collection at each locus Row B) Differences are between single core per-locus measurements. Vertical red lines indicate overall (mean) difference. Black triangles indicate point of no difference (0.0).

Genomic predictive values

The predicative value of each core collection was tested by using it as a training population in a model used to predict traits in the entire collection **(Table 6).** Predictive ability was low for disease score, stand count and flavor across all cores. All cores performed moderately well to predict early season height and disease.

Conclusion:

To incorporate diverse germplasm into breeding programs, researchers need improved strategies for selecting and screening relevant accessions. In this study, we evaluated representative and custom methods of generating core sets of material using 433 accessions of the carrot PI collection from the USDA-NPGS. We found that for this particular crop species representative methods of selecting core sets were equivalently adept at identifying representative sets. Among cultivated carrot accessions, there is only weak genetic substructure. While there is some genetic separation between Eastern and Western breeding pools, the commercial cultivation of carrot around the world implies that geography is not necessarily a good predictor of genetic or phenotypic difference in cultivated accessions. Additionally, phenotypic traits measured in this study vary continuously. Often, discrete morphological traits such as seed color or root shape are used to stratify a collection. If we had access to such data for the PI collection, perhaps more significant differences would have been observed. On the other hand, in the case of a highly admixed population it may simply not be necessary to first stratify a collection before constructing a core set.

For some research goals, a representative set may be what is desired. In other cases, however, the ability to identify non-representative reduced sets of a collection is advantageous. If a desired trait or allele is underrepresented in the collection, a core set that preferentially increases its frequency would be potentially useful. Our custom core sets diverged from the representative sets in terms of phenotypic and genetic diversity. Furthermore, a core set that maintains high correlation between desirable and undesirable traits may be less useful than a non-representative set with accessions that have a lower correlation. In the GBV core used in this study, the correlation in plant height (desirable) and disease score (undesirable) was reduced compared to the total collection. The utility and predictability of these trends needs to be explored further.

Drawing on the tradition of generating a core collection to manage large germplasm resources, in this study we ask how these collections could be designed to be more immediately useful to breeders. Cores that maintain diversity while also maximizing desirable combinations of traits have the potential to be highly valuable to breeders. Future work will evaluate the utility of these custom core strategies to identify and introgression quality and production traits into elite breeding lines.

Acknowledgements

This research was performed in part using the compute resources and assistance of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences.The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. The author thanks the University of Wisconsin Biotechnology Center DNA Sequencing Facility for providing genotyping-by-sequencing library preparation and sequencing services. This work was funded by the National Instituted of Food and Agriculture, USDA (NIFA-USDA) under award numbers 2016-5581-25400 and 2017-67013-26187 and by NIFA-USDA Hatch project 1002863.

Literature Cited

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv:1406.5823 [Stat]*. Retrieved from http://arxiv.org/abs/1406.5823

Beukelaer, H. D., Davenport, G., & Fack, V. (2018). corehunter: Multi-Purpose Core Subset Selection (Version 3.2.1). Retrieved from https://CRAN.R-project.org/package=corehunter

Boiteux, L. S., Fonseca, M. E. N., & Simon, P. W. (1999). Effects of Plant Tissue and DNA Purification Method on Randomly Amplified Polymorphic DNA-based Genetic Fingerprinting Analysis in Carrot. *Journal of the American Society for Horticultural Science*, *124*(1), 32–38.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics (Oxford, England)*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Brown, A. (1989). Core collections: A practical approach to genetic resource management. *Genome*, *31*, 818–824. https://doi.org/10.1139/g89-144

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, *4*(3), 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Ersts, P. (n.d.). Geographic Distance Matrix Generator. American Museum of Natural History, Center for Biodiversity and Conservation. (Version 1.2.3). American Museum of Natural History, Center for Biodiversity and Conservation. Retrieved from http://biodiversityinformatics.amnh.org/open_source/gdmg.

FactoMineR.pdf. (n.d.). Retrieved from https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLOS ONE*, *9*(2), e90346. https://doi.org/10.1371/journal.pone.0090346

hierfstat, a package for r to compute and test hierarchical F-statistics - GOUDET - 2005 - Molecular Ecology Notes - Wiley Online Library. (n.d.). Retrieved April 22, 2018, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-8286.2004.00828.x

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., … Simon, P. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics*, *48*(6), 657–666. https://doi.org/10.1038/ng.3565

Luby, C. H., Dawson, J. C., & Goldman, I. L. (2016). Assessment and Accessibility of Phenotypic and Genotypic Diversity of Carrot (Daucus carota L. var. sativus) Cultivars Commercially Available in the United States. *PLOS ONE*, *11*(12), e0167865. https://doi.org/10.1371/journal.pone.0167865

Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, *8*(19), 4321–4325.

Noirot, M., Hamon, S., & Anthony, F. (1996). The principal component scoring: A new method of constituting a core collection using quantitative data. *Genetic Resources and Crop Evolution*, *43*(1), 1–6. https://doi.org/10.1007/BF00126934

Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. L. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *126*(2), 289–305. https://doi.org/10.1007/s00122-012-1971-y

Oliveira, M. F., Nelson, R. L., Geraldi, I. O., Cruz, C. D., & de Toledo, J. F. F. (2010). Establishing a soybean germplasm core collection. *Field Crops Research*, *119*(2–3), 277–289. https://doi.org/10.1016/j.fcr.2010.07.021

R: The R Stats Package. (n.d.). Retrieved April 22, 2018, from https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html

Skroch, P. W., Nienhuis, J., Beebe, S., Tohme, J., & Pedraza, F. (1998). Comparison of Mexican common bean (Phaseolus vulgaris L.) core and reserve germplasm collections. *Crop Science*, *38*(2), 488–496.

TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. (n.d.). Retrieved April 22, 2018, from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090346

Thachuk, C., Crossa, J., Franco, J., Dreisigacker, S., Warburton, M., & Davenport, G. F. (2009). Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. *BMC Bioinformatics*, *10*, 243. https://doi.org/10.1186/1471-2105-10-243

Theisen, T. (2016). *Organic Open Pollinated Carrot Phenotyping: Returning to the Roots for Improving Organic Production in Main and Cold Season Cultivation*. University of Wisconsin- Madison, Madison, Wisconsin.

Wickham, H., Chang, W., & RStudio. (2016). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics (Version 2.2.1). Retrieved from https://CRAN.R-project.org/package=ggplot2

Sup. Fig. S3

Genotype and SNP filtering workflow. SNPs were filtered into two datasets, D1 and D2. D1 was further filtered to excluded samples from the Portuguese-W STRUCTURE group (D1-noPT). D2 was further filtered to exclude samples with more than 30% admixture as determined by STRUCTURE (D2-lowAd). Gray shaded boxes show the number of SNPs and individuals in each dataset. GW; Genome-wide, SW; Sliding Window.

Sup. Fig. S4 Genome-wide linkage disequilibrium (r^2) in wild (black triangles) and domesticated (orange triangles) carrots. LD decay rate is represented by the intersection of the fitted LD decay curve with r^2 = 0.1 and r^2 = 0.2.

Sup. Fig. S5 Plot of the Δ K to determine the most likely substructuring of carrot accessions based on STRUCTURE.

Sup. Fig. S6

Geographic distribution and population structure of 674 carrot accessions (D2). A) Geographic distribution of accessions each represented by a point on the map colored according to STRUCTURE group. Commercial cultivated samples not shown. B) PCA plot of the first two principal components. PC1 and PC2 account for 12.4% and 4.6% of total variation, respectively. C) STRUCTURE groups. Percentage of membership (q) for each group as identified at K=6. Population structure analysis using STRUCTURE. Each color represents a single population. Each vertical column represents one accession and each colored segment in each column represents the proportion contributed from ancestral populations. The 674 accessions were divided into six groups. D) Maximum-likelihood tree of carrot accessions with the outgroup *Daucus syrticus* shown in black. Numbers on the branches indicate bootstrap support. E) Color and sample key based on K = 6.

Sup. Fig. S7

Population structure analysis of wild and domesticated carrot accessions at all K between 2 and 6. Geographic groupings are listed as well as cultivation status (W; Wild, D; Domesticated, HI; Hybrid Imperator).

Sup. Fig. S8

Population structure analysis of wild and domesticated carrot accessions at all K between 2 and 6 using D2-lowAd. Geographic groupings are listed as well as cultivation status (W; Wild, D; Domesticated, HI; Hybrid Imperator).

Sup. Fig. S9 QQ plot for GWAS analysis for orange carrot root color.

Sup. Fig S10 Genome-wide nucleotide diversity (π) in wild and domesticated carrot accessions. Sliding window analysis of 500 kb regions plotting nucleotide diversity difference between wild and domesticated carrot accessions. Red line indicates the top 5% of values.

Sup. Fig. S11 Genome-wide F_{st} between wild and domesticated carrot accessions. Sliding window analysis of 500 kb regions plotting F_{st} between wild and domesticated carrot accessions. Red line indicates the top 5% of values.

Sup. Fig. S12 Genome-wide XP-CLR between wild and domesticated carrot accessions. Sliding window analysis of averaged 10 kb regions plotting XP-CLR (wild as reference population and domesticated carrot accessions as object population). Red line indicates the top 1% of values.

Sup. Fig. S13 Genome-wide linkage disequilibrium averaged across sliding windows of 100 SNPs in wild carrots. Regions identified as significant in Figure 4 A and B are highlighted in orange.

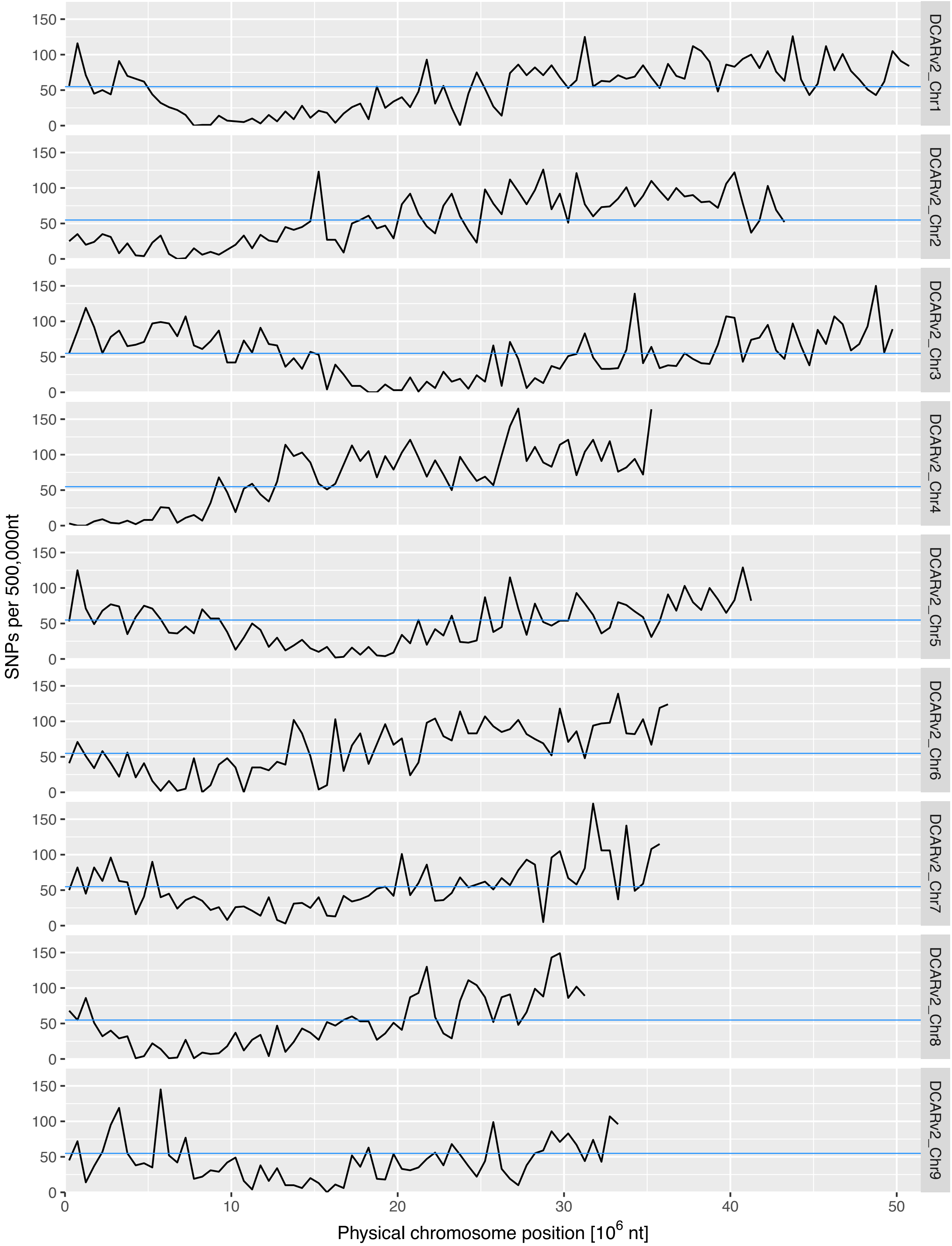


Figure S1 Average SNP density in 500 kb bins across the nine chromosomes for D1. Blue line is genome wide average of 54 SNPs/500 kb.

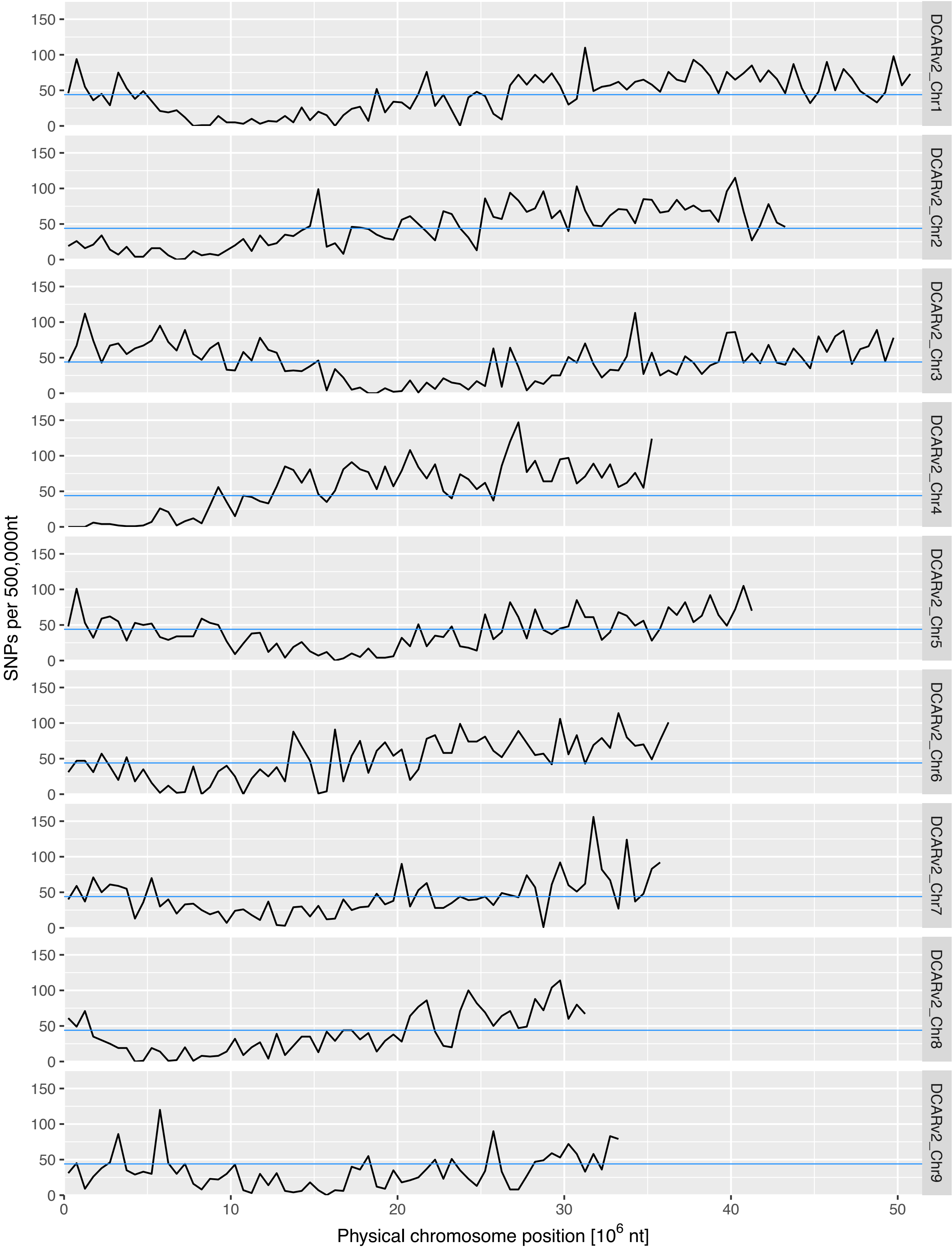


Figure S2 Average SNP density in 500 kb bins across the nine chromosomes for D2. Blue line is genome wide average of 43 SNPs/500 kb.

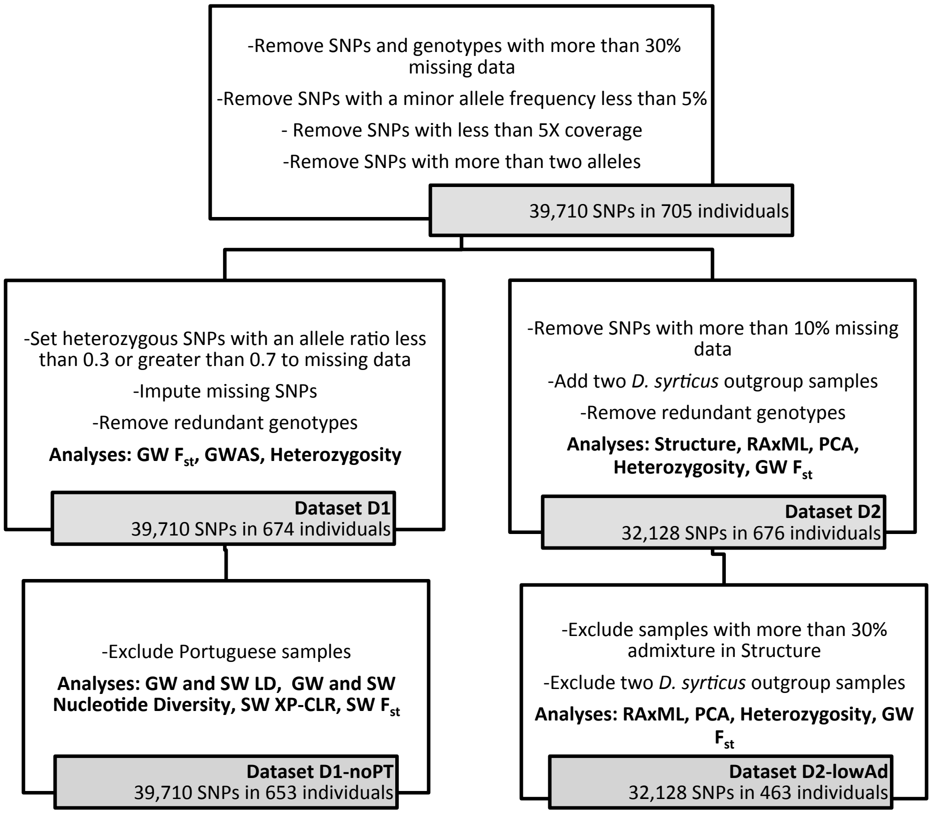

-Remove SNPs and genotypes with more than 30% missing data
-Remove SNPs with a minor allele frequency less than 5%
- Remove SNPs with less than 5X coverage
-Remove SNPs with more than two alleles
39,710 SNPs in 705 individuals
-Set heterozygous SNPs with an allele ratio less than 0.3 or greater than 0.7 to missing data
-Impute missing SNPs
-Remove redundant genotypes
Analyses: GW F$_{st}$, GWAS, Heterozygosity
Dataset D1
39,710 SNPs in 674 individuals
-Remove SNPs with more than 10% missing data
-Add two *D. syrticus* outgroup samples
-Remove redundant genotypes
Analyses: Structure, RAxML, PCA, Heterozygosity, GW F$_{st}$
Dataset D2
32,128 SNPs in 676 individuals
-Exclude Portuguese samples
Analyses: GW and SW LD, GW and SW Nucleotide Diversity, SW XP-CLR, SW F$_{st}$
Dataset D1-noPT
39,710 SNPs in 653 individuals
-Exclude samples with more than 30% admixture in Structure
-Exclude two *D. syrticus* outgroup samples
Analyses: RAxML, PCA, Heterozygosity, GW F$_{st}$
Dataset D2-lowAd
32,128 SNPs in 463 individuals

Figure S3 Genotype and SNP filtering workflow. SNPs were filtered into two datasets, D1 and D2. D1 was further filtered to excluded samples from the Portuguese-W STRUCTURE group (D1-noPT). D2 was further filtered to exclude samples with more than 30% admixture as determined by STRUCTURE (D2-lowAd). Gray shaded boxes show the number of SNPs and individuals in each dataset. GW; Genome-wide, SW; Sliding Window.

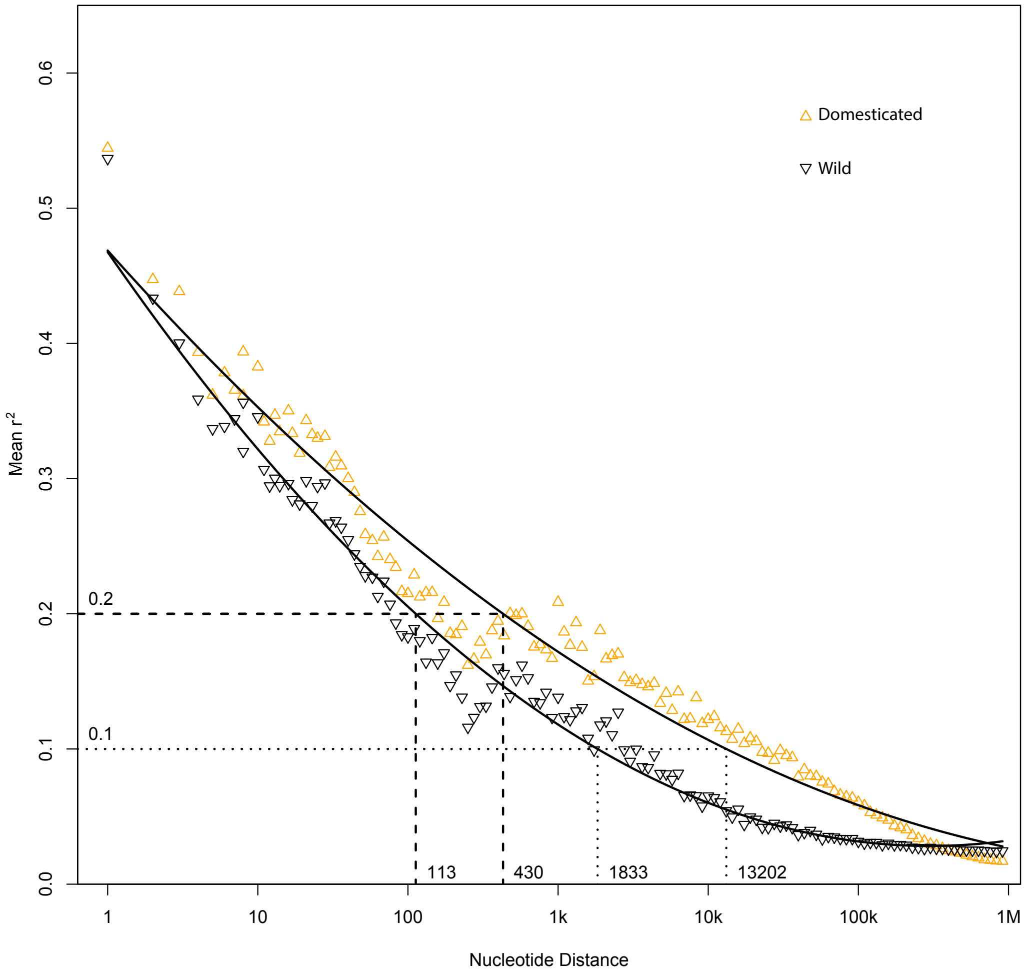


Figure S4 Genome-wide linkage disequilibrium (r^2) in wild (black triangles) and domesticated (orange triangles) carrots. LD decay rate is represented by the intersection of the fitted LD decay curve with $r^2 = 0.1$ and $r^2 = 0.2$.

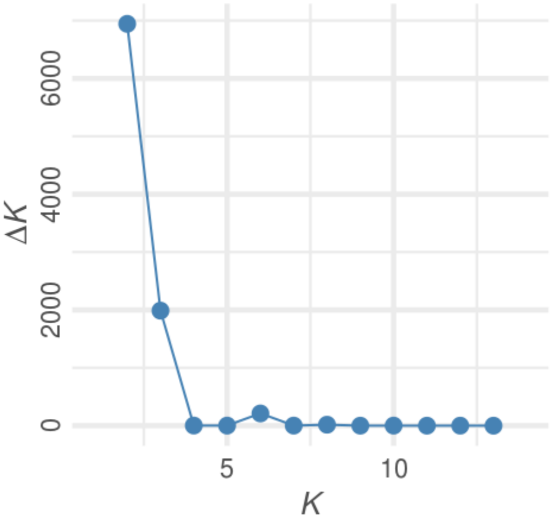


Figure S5 Plot of the Δ K to determine the most likely substructuring of carrot accessions based on STRUCTURE.

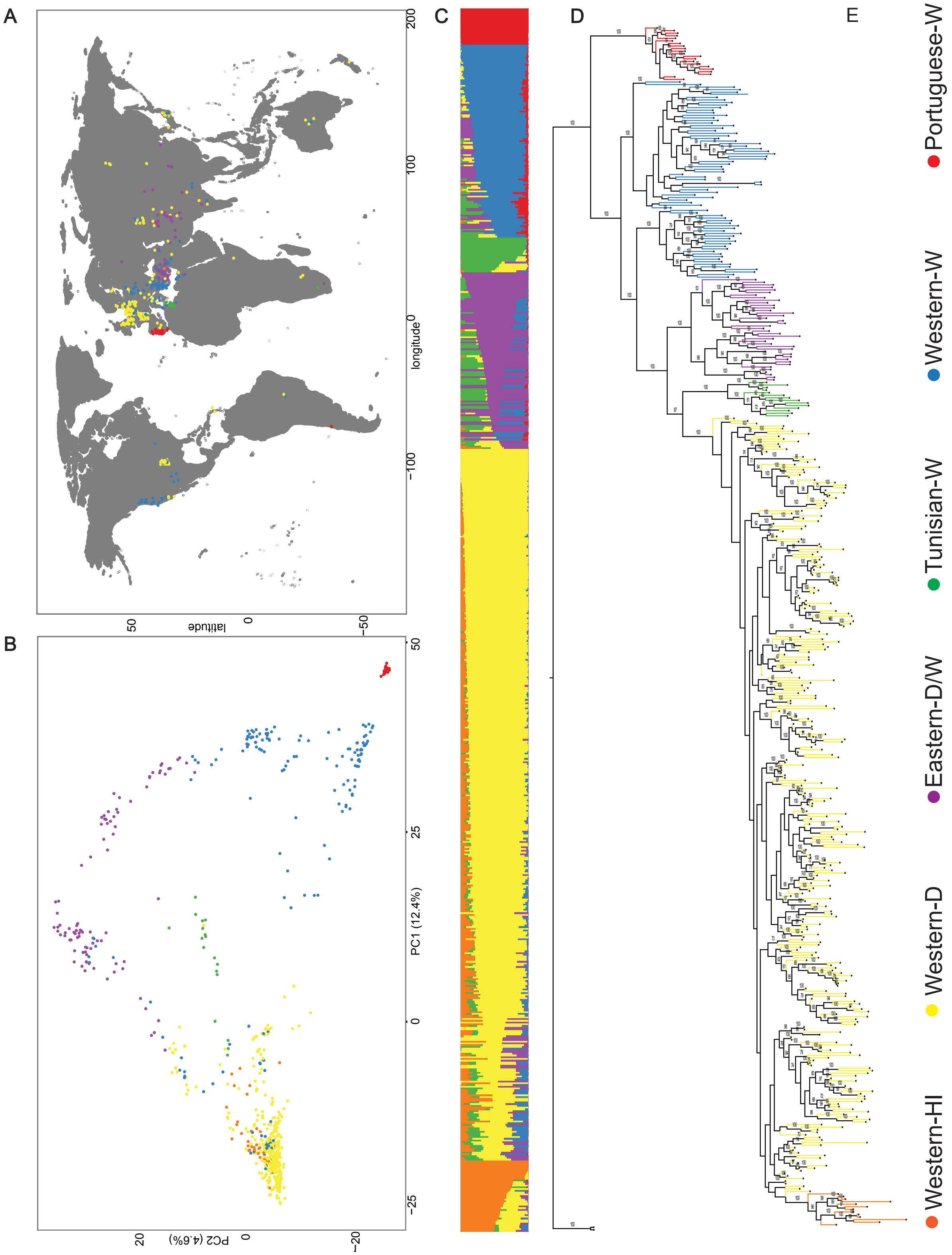


Figure S6 Geographic distribution and population structure of 674 carrot accessions (D2). A) Geographic distribution of accessions each represented by a point on the map colored according to STRUCTURE group. Commercial cultivated samples not shown. B) PCA plot of the first two principal components. PC1 and PC2 account for 12.4% and 4.6% of total variation, respectively. C) STRUCTURE groups. Percentage of membership (q) for each group as identified at K=6. Population structure analysis using STRUCTURE. Each color represents a single population. Each vertical column represents one accession and each colored segment in each column represents the proportion contributed from ancestral populations. The 674 accessions were divided into six groups. D) Maximum-likelihood tree of carrot accessions with the outgroup *Daucus syrticus* shown in black. Numbers on the branches indicate bootstrap support. E) Color and sample key based on K = 6.

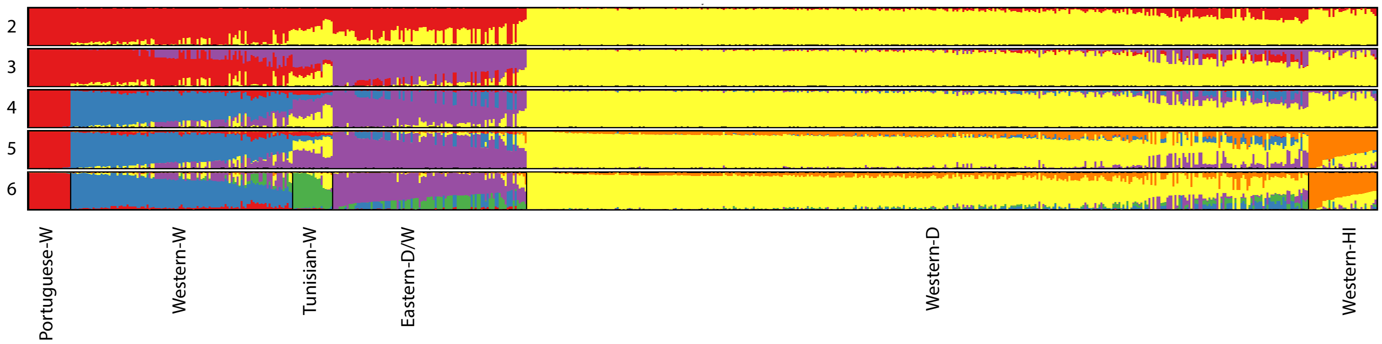


Figure S7 Population structure analysis of wild and domesticated carrot accessions at all K between 2 and 6. Geographic groupings are listed as well as cultivation status (W; Wild, D; Domesticated, HI; Hybrid Imperator).

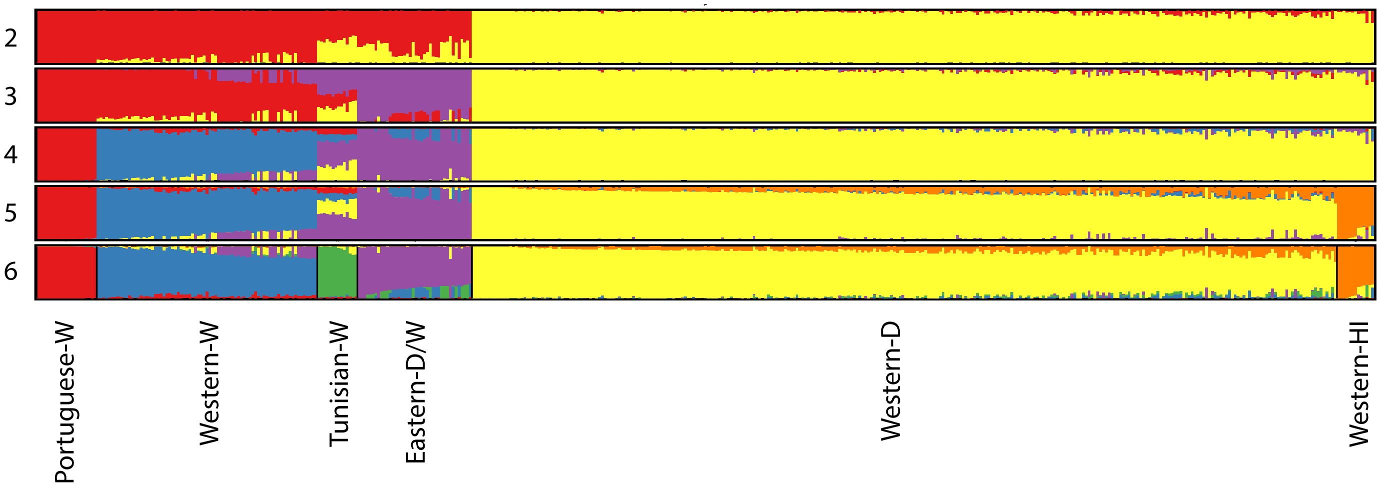


Figure S8 Population structure analysis of wild and domesticated carrot accessions at all K between 2 and 6 using D2-lowAd. Geographic groupings are listed as well as cultivation status (W; Wild, D; Domesticated, HI; Hybrid Imperator).

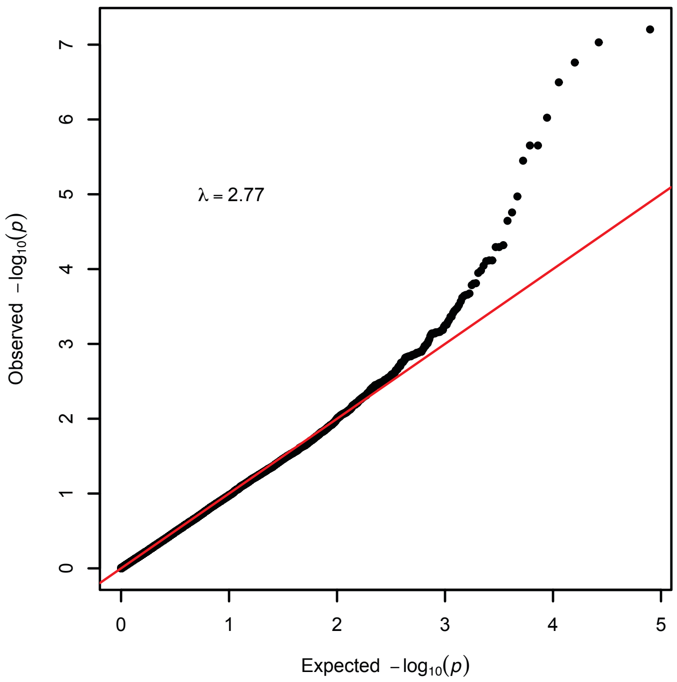


Figure S9 QQ plot for GWAS analysis for orange carrot root color.

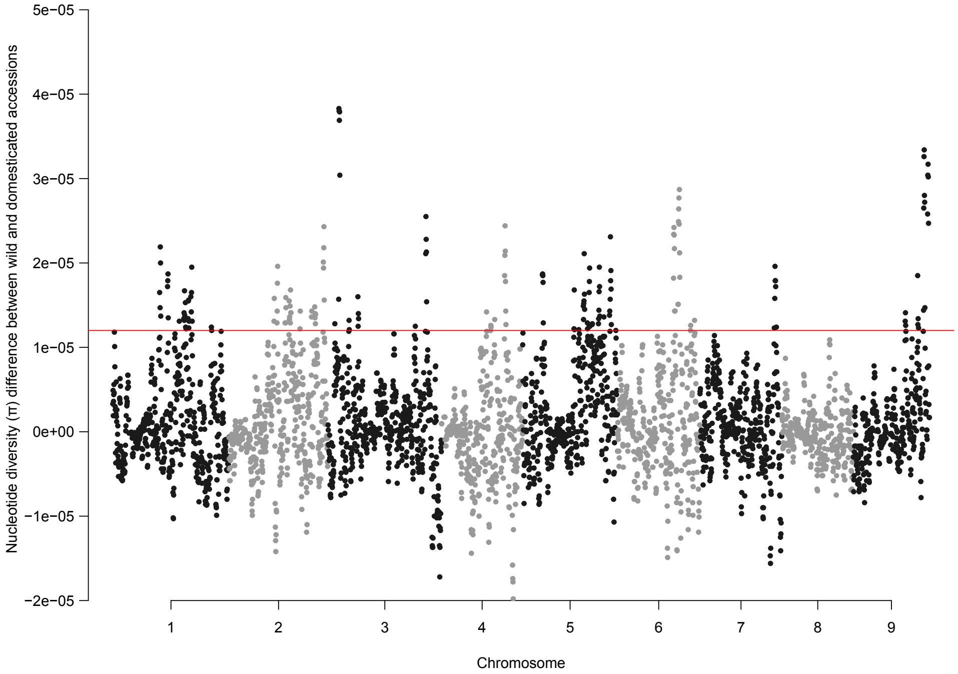


Figure S10 Genome-wide nucleotide diversity (π) in wild and domesticated carrot accessions. Sliding window analysis of 500 kb regions plotting nucleotide diversity difference between wild and domesticated carrot accessions. Red line indicates the top 5% of values.

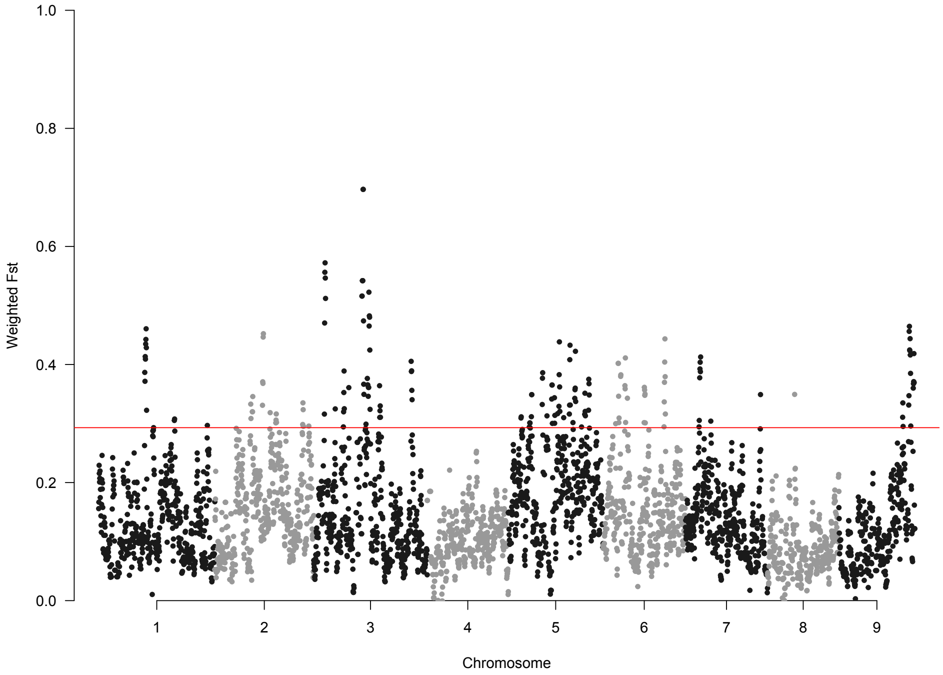


Figure S11 Genome-wide F_{st} between wild and domesticated carrot accessions. 500 kb regions that are likely to contain a selective sweep by appearing in the top 5% of all three tests (nucleotide diversity, F_{st}, and XP-CLR) are shown in red.

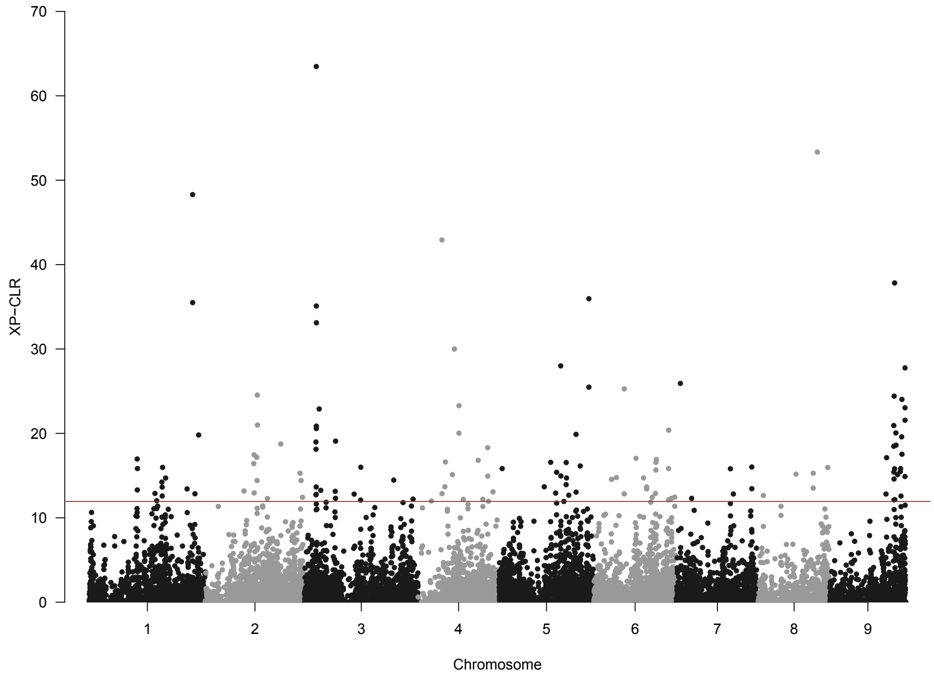


Figure S12 Genome-wide XP-CLR between wild and domesticated carrot accessions. Sliding window analysis of averaged 10 kb regions plotting XP-CLR (wild as reference population and domesticated carrot accessions as object population). Red line indicates the top 1% of values.

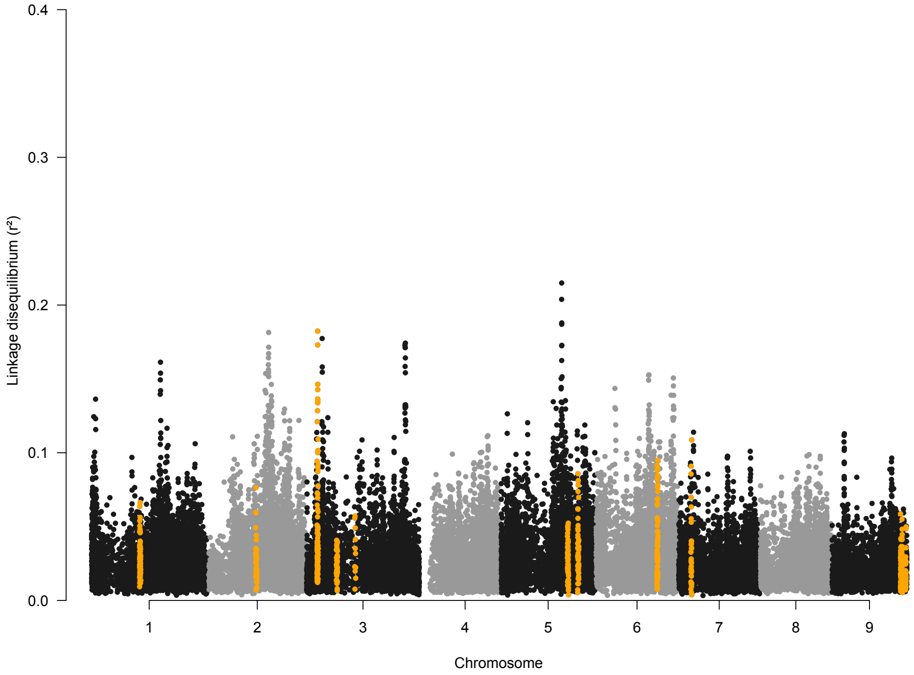


Figure S13 Genome-wide linkage disequilibrium averaged across sliding windows of 100 SNPs in wild carrots. Regions identified as significant in Figure 4 A and B are highlighted in orange.

Appendix B: Supplementary Table 1 (Ch 3): Passport data and LS-mean estimates phenotypes for 433 PIs

PI	Name	Origin	est.latitude	est.longitude	early_height_1	early_height_2	early_height_3	early_width_1	early_width_2	early_width_3	late_height_1	late_height_2	late_height_3	disease_score	late_width_1	late_width_2	late_width_3	stand_count	harshness	sweetness
Ames_17826	Estonia	Estonia	58.595	25.014	15.51	12.69	16.55	21.5	21.98	20.76	24.19	32.25	27.18	3.17	NA	NA	NA	NA	NA	NA
Ames_22389	Long Red	Nepal	28.395	84.124	17.24	16.56	16.2	50.95	38.57	37.11	44.04	70.6	64.32	2.33	60.75	62.6	62.78	4	NA	NA
Ames_27398	Z019	Uzbekistan	41.377	64.585	28.63	26.63	27.25	28.75	32.88	31.63	54	72	51.37	3.23	52.5	54.5	53.22	5	2.5	3
Ames_27399	Z020	Uzbekistan	41.377	64.585	28.32	23.36	19	42.68	41.98	35.3	51.82	78.47	54.56	2.89	40.25	40.4	37.22	3	NA	NA
Ames_27400	Z021	Uzbekistan	41.377	64.585	23	23.38	23	28.63	32.5	37.75	45.25	60.5	44	3.23	44.5	55	46	3.5	2.75	3.5
Ames_27414	Z109	Uzbekistan	41.377	64.585	15.65	15.19	15.33	24.68	25.14	22.3	33.65	26.8	36.5	3.23	35.25	37.4	36.22	2	2	2.83
Ames_29084	19	Tunisia	33.887	9.537	23.25	28.88	21.5	21.99	20.84	32.55	39.97	50.98	40.68	3.23	NA	NA	NA	5.5	NA	NA
Ames_29182	GSMO 2-28	Georgia,_South_Ossetia	42.168	44.377	21.65	24.36	22.66	22.02	26.14	23.3	46.82	59.47	46.56	3.23	46.25	48.4	51.22	3.5	2.83	2.5
Ames_30276	Tun340	Tunisia,_Kairouan	35.676	10.092	26.35	21.6	23.03	26.65	33.34	31.05	39.47	125.06	37.75	3.23	22.75	NA	NA	2	NA	NA
Ames_7701	Juwarot	Germany,_Saxony-Anhalt	51.95	11.692	12.75	10	13.87	12.13	12.81	14.3	31.75	35.25	32.62	2.56	23.5	17	NA	1	2.33	3.33
Ames_7702	Nantskaja Char kovskaja	Russian_Federation	61.524	105.319	19	21.75	19.62	21.5	24.81	23.13	36.5	45.75	34.5	3.56	36.5	35.5	33.22	9	3	3.5
Ames_7705	Chibinskaja	Russian_Federation	61.524	105.319	23.99	26.69	28	35.02	36.98	37.63	35.99	56.8	37.39	3.56	42.25	44.4	55.22	7	3	2.67
Ames_7711	Fortuna	Germany,_Saxony-Anhalt	51.95	11.692	21.25	22.38	22.25	30.75	26.25	26.13	48	73.25	49.37	3.56	53	50.5	50.5	11.5	NA	NA
Ames_7715	Amager	Poland	51.919	19.145	24	22.38	21.87	32	29.38	26.25	48.12	68.5	42.25	2.56	58	56	58	3.5	3	3
NSL_199857	F524C	United_States,_Wisconsin	43.784	-88.788	11.5	10.88	9.37	13.88	14.63	18.25	23.75	42.5	26.25	2.56	45	38.5	32.5	3.5	4	2.5
NSL_199859	6439M	United_States,_Wisconsin	43.784	-88.788	14	15.75	13.12	14.88	18.25	16.25	34.62	38.5	33	3.23	34.5	33.5	28.5	2.5	3.5	2.83
NSL_199860	6274M	United_States,_Wisconsin	43.784	-88.788	11.26	12.44	12.8	16	15.48	17.26	21.94	8.25	21.18	2.89	NA	NA	NA	0	3.17	3.83
NSL_199861	4367S	United_States,_Wisconsin	43.784	-88.788	18.25	20.13	17.25	24	28.5	26	35	49	39.5	2.89	48.5	42	51.5	3.5	1.83	3
NSL_199865	3080M	United_States,_Wisconsin	43.784	-88.788	15.25	14.25	14.87	19.75	18	19.13	34.5	45	31.5	3.56	39.5	32.5	38.5	7.5	1.17	2.83
NSL_199868	9253M	United_States,_Wisconsin	43.784	-88.788	12.25	11.5	12.62	18.75	17.38	14.25	22.87	26.5	21	3.23	37	36.5	32	2.5	3.17	3
NSL_26501	EMPRESS	United_States,_Connecticut	41.603	-73.088	19	16.63	19.12	16.75	19.75	21.13	45.5	54.75	38	3.23	33.5	32	35.5	3.5	3.33	2.83
NSL_26502	LONG IMPERATOR 11B	United_States,_Michigan	44.315	-85.602	20.88	19	20.62	20.5	24.88	23	45.12	58.5	40.75	3.23	55	37.5	35	3	3.5	3
NSL_34344	WISSYN 6	United_States,_Wisconsin	43.784	-88.788	18.88	18.75	17.5	15	17.25	14.75	39	48.75	37.62	3.23	26	35	25.5	11	2.25	3
NSL_34346	WISSYN 171	United_States,_Wisconsin	43.784	-88.788	10.88	13.63	13	11.13	14	14.63	26.12	37.25	28.75	2.89	31.5	23.5	27	3.5	2.5	3
NSL_52533	PACESETTER	United_States,_Minnesota	46.73	-94.686	19.5	21.13	18.37	21.75	22.25	21.75	39.75	46.25	41	3.56	38	40.5	40	6.5	4	2.67
NSL_54098	HICOLOR 9	United_States,_Michigan	44.315	-85.602	23.13	27	23	28.75	32.75	38.25	43.62	73.5	48.87	3.23	73.25	61.4	65.22	7.5	3	2.83
NSL_6166	CHANTENAY RED CORE H	United_States,_California	36.778	-119.418	26.5	28.25	28.87	27.75	25.38	30	47.87	67.25	46.87	3.56	56.5	59.5	58	8.5	2.83	3.17
NSL_6168	CHANTENAY ROYAL	United_States,_Colorado	39.55	-105.782	33.63	25.5	35.62	34.75	36.5	41.75	58.75	89.75	55.37	3.56	77	70.5	75	18.5	2	3.25
NSL_6172	DANVERS RED CORE	United_States,_California	36.778	-119.418	35.75	34.75	32.37	43.5	43.25	39.25	64.87	94.25	63	2.89	61	74.5	60.5	15	3.75	3
NSL_65838	GOLD PAK 28	United_States,_California	36.778	-119.418	21.88	20.25	21.5	27.13	24.63	32.75	51	67	44.12	3.56	57	47	57.5	7.5	3.67	2.67
NSL_9333	NANTES CORELESS	United_States,_Minnesota	46.73	-94.686	23.13	26.5	22.12	24.25	18.25	23.5	44.37	62.75	49.5	3.56	45.5	43.5	38.5	9	2.33	3
PI_163234	Gajar	India,_Madhya_Pradesh	22.973	78.657	18.25	22.75	16.75	28.13	32.38	23.25	40.87	50.75	45.87	3.23	46	49	47.5	1.5	3.75	1.75
PI_163235	Gajar	Pakistan,_Punjab	31.17	72.71	23.25	23	99.37	33.75	32.75	34	42.9	67.52	43.68	3.56	NA	NA	NA	12	4	4
PI_164136	Gajar	India,_Madhya_Pradesh	22.973	78.657	29.38	30	32.62	28.5	32.75	31	62.47	105.37	60.23	2.89	50.75	47.6	73.78	9	2.83	3.5
PI_164461	Gajar	India,_Rajasthan	27.024	74.218	25.38	23.5	22.87	25	24.25	28	49.5	76	52.25	3.23	63	62	60	11.5	3.33	2
PI_164484	Gajar	India,_Rajasthan	27.024	74.218	31.49	30.03	33.16	34.02	34.14	36.63	61.49	93.13	60.73	3.56	63.25	75.4	59.22	4	3.75	3
PI_164942	Kartal	Turkey,_Istanbul	41.008	28.978	20.32	21.5	24.75	30.25	28.5	30	50.5	68.5	53	3.23	54.5	48.5	37	4	2.17	3
PI_164943	19	Turkey,_Istanbul	41.008	28.978	24.75	23.75	23.5	30.63	34.25	32.5	45	64.5	46	2.89	70.5	67.5	64.5	4	2.17	2.33
PI_165484	Gajar	India,_Uttar_Pradesh	26.847	80.946	20.88	14.75	13.32	21	24.5	24.07	36.62	47.5	43.44	2.89	30.5	34.5	35.5	1.5	3.5	3.5
PI_165522	Gajar	India	20.594	78.963	19.13	19	17.25	16.88	21	16	41.25	59	37.62	2.89	30.5	33.5	30.5	2	3	3.5
PI_167143	340	Turkey,_Icel	36.812	34.641	28.25	30.5	30	42.5	50	41.75	47.19	72.25	48.18	3.23	NA	NA	NA	12.5	1.67	3.17
PI_167211	Havuc	Turkey,_Icel	36.812	34.641	26.75	25.25	27.87	33.99	33.18	32.05	50.44	60.75	41.43	3.89	NA	NA	NA	16	2	3.17
PI_169480	1839	Turkey,_Mugla	37.215	28.363	32.75	32.63	32.62	35.25	27.75	30.88	56.5	82.5	54.62	3.89	63.5	76	72.5	8	2.83	3
PI_169482	2198	Turkey,_Manisa	38.614	27.43	26	26.75	28.87	26.75	25.63	20.5	45	65.5	49.37	3.23	48	48.5	46	5	2	3
PI_169483	2231	Turkey,_Izmir	38.424	27.143	18.63	20.5	18.75	30.5	36.75	31.75	42.62	57.25	42.25	3.56	44	48	53	9	1.33	3
PI_169486	2625	Turkey,_Kirklareli	41.735	27.224	16.13	19	19.87	24.75	25.25	25	40.12	67.5	43.62	3.23	42	49	49.5	4.5	2.33	3.33
PI_169487	2701	Turkey,_Edirne	41.677	26.556	25.63	25.5	23	29	33	30.25	48.25	74.25	48.62	2.89	43	47.5	47.5	12.5	2.83	3.33
PI_169490	3583	Turkey,_Bilecik	40.143	29.979	20.38	19.63	19	20.75	19.75	17.88	44.37	64	45.25	3.56	38	41.5	38	6.5	3	3.17
PI_171641	6821	Turkey,_Tokat	40.323	36.552	18.63	20.75	20	18.75	23	22.25	52.37	77	50.5	3.23	39	40	35	3	3.33	3.67
PI_171645	7306	Turkey,_Erzurum	39.905	41.266	15.52	14.27	18.37	17.15	22.18	18.38	41.3	58.03	47.23	3.23	40.75	39.6	26.78	3	2.5	3
PI_172886	7518	Turkey,_Kars	40.601	43.097	18.38	16.25	13.37	27.5	25.75	20.63	39.75	47.5	39.12	3.23	36	40.5	34.5	8	2.67	3.5
PI_172893	8496	Turkey,_Maras	37.575	36.923	20.02	16.6	17.2	21.32	25.34	32.01	32.47	31.37	21.5	2.46	10.75	24.6	21.78	0	2.83	2.67
PI_173687	7882	Turkey,_Hakkari	37.577	43.737	17.13	20.25	19.25	28.5	24.75	22	39.75	48.5	44.75	3.56	33	40	51	8.5	2	3
PI_173688	8612	Turkey,_Malatya	38.355	38.334	16.69	16.44	10.7	17.32	24.18	28.01	42.14	51.7	41.89	2.89	55.75	27.6	NA	5.5	2	3
PI_174202	8073	Turkey,_Diyarbakir	37.925	40.211	22.38	21.88	25.87	25.5	25.5	26.75	40.5	74.75	37	3.56	24	25	28.5	17.5	2.83	3.5
PI_174205	8123	Turkey,_Mardin	37.313	40.734	24.25	20.25	21.37	25.25	22	21	41.37	67.25	40	2.89	33.5	30	33	11	2.5	3.17
PI_174206	8238	Turkey,_Urfa	37.167	38.796	23.88	23.13	24	32.5	22	20.5	38.25	81.5	40.75	3.23	27.5	25.6	20.78	7.5	2.17	3
PI_174208	8791	Turkey,_Elazig	38.675	39.223	22.65	27.69	28	32.02	34.48	32.13	51.94	71.75	40.18	3.56	NA	NA	NA	3.5	2.5	2.5
PI_174828	Gajar	India,_Uttar_Pradesh	26.847	80.946	29.13	27.25	27.5	41.75	37.5	40.75	51.87	77.75	56	3.89	50.5	49.5	62	9	3.17	2.83
PI_175719	9714	Turkey,_Eskisehir	39.767	30.526	20.88	17.25	22.87	32.75	21.25	25.5	41.87	60.25	51	3.23	44.5	36.5	36	2	2.67	3
PI_176556	8848	Turkey,_Erzincan	39.747	39.491	17.75	18.63	18.87	31.5	31	30.25	47.5	72.75	49.37	3.56	33.5	41.5	40	14	2.17	2.67
PI_176557	8990	Turkey,_Sivas	39.751	37.015	20.75	20.25	19.87	30.75	22.5	29	55.87	68	49.75	3.23	41.5	33	49.5	7.5	2.17	3
PI_176561	9426	Turkey,_Afyon	38.757	30.539	21.88	22	24.25	26.75	27.75	30.5	48.25	73.5	46.75	3.56	25.5	46	30	8	2.17	3.17
PI_176563	9482	Turkey,_Kutahya	39.42	29.986	12	11.63	9.65	13.25	12.88	8.57	31.25	29.25	38.11	3.04	26	25	23	1.5	NA	NA
PI_176565	9541	Turkey,_Bilecik	40.143	29.979	28.88	34.13	27.5	26.25	33.75	31.75	54.8	83.37	55.23	3.89	71.75	66.6	55.78	11.5	2.5	3
PI_176970	9359	Turkey,_Konya	37.875	32.493	26.13	26.88	25	26.25	33.75	29.25	45.75	45.25	45.25	3.04	32.5	33.5	45.5	6.5	1.83	3
PI_177384	Beledi	Syria	34.802	38.997	16.88	17.63	17.62	30.25	26	23.38	23.12	58.75	23.89	3.04	16.5	11.6	NA	6.5	2.25	2.75
PI_179275	4966	Turkey,_Corum	40.55	34.954	23.75	22.5	21.25	30.75	33	30.5	50.5	68.75	47	3.23	56	62	62.5	12	2	2.75
PI_180834	5012	Turkey,_Zonguldak	41.454	31.789	32.13	26	26.5	28.75	35	22.5	57.5	93.25	61	3.23	45	55	50.5	9	2	3.5
PI_181052	8604	Pakistan,_Sind	25.894	68.525	20.25	22	19.87	23	22.5	21.75	49.62	67.25	36.37	2.56	47	47	31	2.5	2.67	3
PI_181765	9949	Lebanon	33.855	35.862	24.5	21.75	23	21.25	32.25	31	43.5	61.75	48.5	2.89	58.5	65	61	10	2.33	2.67
PI_181767	9973	Lebanon	33.855	35.862	15.63	15.38	13.62	23.25	23.25	19.75	31.3	46.03	32.73	3.23	54.75	50.6	45.78	7.5	2.17	2.67
PI_182204	10449	Turkey,_Gumushane	40.461	39.48	14	10.38	13	23.75	19	22.5	37.99	44.8	34.56	3.23	41.25	32.4	39.22	2	1.83	3
PI_182206	10587	Turkey,_Bitlis	38.401	42.11	17.63	15.88	11	13.25	15.75	8.25	40.37	49.75	35.37	3.23	25	29	22	0	3	3.5
PI_187234	Red Giant (Obtuse of Fla	Belgium	50.504	4.47	15.5	11.13	13.75	13.75	13.25	11.38	34.87	36.75	31.75	2.56	45.5	34	27.5	2.5	2.33	3.17
PI_187235	Nantes No. 1	Belgium	50.504	4.47	22.88	21.25	21.25	16.5	20.25	20.13	31.75	43	36.62	3.56	35	51.5	49	5	3.5	3.33
PI_187236	Nantes No.2	Belgium	50.504	4.47	20.88	18.88	20.25	17	17	19.75	35.25	40.5	37.5	3.56	36	38.5	38.5	4	3.5	2.83
PI_187237	Touchon	Belgium	50.504	4.47	14	17.25	19.25	15.25	16	17.13	32.25	43.75	34.75	3.23	27.5	24.5	31	7.5	2.83	3
PI_193504	Nantaise	Ethiopia	9.145	40.49	16.63	19.63	21	18	24.5	19.75	42.87	61.5	38.75	2.89	32	38.5	33	6.5	2.33	2.83
PI_196847	10065	Ethiopia	9.145	40.49	20.75	19.38	20.37	21.75	27.5	28	38.87	49.75	38.75	3.23	43.5	51.5	48.5	12	3	3
PI_200876	RWL 4275	Afghanistan	33.939	67.71	3.94	1.69	-0.6	19.04	5.47	-0.14	21.59	25.9	20.32	2.33	24.25	37.4	12.22	3	NA	NA
PI_204704	426	Turkey,_Malatya	38.355	38.334	26.63	32.5	29.87	41.5	38.75	34.25	49	77.75	48.75	3.56	52	50.5	57.5	11	2.67	3
PI_205999	Regulus W:s/44	Sweden	60.128	18.644	17	18.13	19.37	21.63	24.5	21.5	46	54.5	47.75	2.39	50	42	52	11.5	2.5	2.83
PI_211024	12977	Afghanistan,_Herat	34.353	62.204	30	24.63	21.75	34	33	38.5	38.87	44.75	38.56	3.23	35	47.5	45.22	11	1.83	3
PI_211590	12770	Afghanistan,_Badakhshan	36.735	70.812	13.5	13.63	11.5	23.5	26.5	20	34.75	52.75	35.62	3.23	41	36	34.78	1.5	2.67	3
PI_220014	Zardak (Carrot)	Afghanistan,_Kabul	34.555	69.207	20.38	16.25	18.87	24.5	29.75	30.75	36.14	37.7	35.06	3.89	55.75	50.6	33.78	16	2.5	3
PI_220657	Zardak (Carrot)	Afghanistan,_Herat	34.353	62.204	28.63	28.88	28.25	38.75	36	35.5	50.75	72	48	3.89	75.5	66	67	10.5	2.5	3
PI_220795	450	Afghanistan,_Kondoz	36.729	68.868	26.75	24.63	25.5	30.25	33	34.25	51.25	79	49.12	3.89	50	63	63	8	1.67	2.83
PI_221924	Zardak (Carrot)	Afghanistan,_Paktia	33.706	69.383	24.25	25.75	21.5	29.13	28.75	28.25	39	38	38.62	3.73	46.5	36	42.5	13.5	2.17	3.83
PI_222249	1436	Iran,_Tehran	35.689	51.389	16.5	15.13	13.5	16.88	25	17.5	32.87	36.25	23.37	2.89	46.5	38	26.5	1.5	NA	NA
PI_222250	1437	Iran,_Tehran	35.689	51.389	25.63	25.38	26.37	29.25	39	31	41.19	59.75	44.18	3.89	NA	NA	NA	13	2.67	2.83
PI_223361	1540	Iran	32.428	53.688	17	16.63	17.5	29	24	26.63	45.75	80.25	39.87	3.23	35.5	45	36.5	2	2.75	3
PI_224689		Myanmar	21.916	95.956	22.13	22.38	20.75	27	25	29.5	43.5	54	39.75	3.56	51	38	44.5	5.5	2.83	3.67
PI_225866	Amager No. 23	Denmark	56.264	9.502	21.38	21.25	21	21.25	21	20.3	40.87	55.75	38.75	2.89	45.5	50.5	35	1.5	2.67	3
PI_225867	Amsterdam No. 378	Denmark	56.264	9.502	14.99	17.03	19	16.35	13.48	16.63	28.82	23.13	27.73	3.23	36.25	34.4	29.22	1.5	2.67	3
PI_225868	Chantenay Red Core No.	Denmark	56.264	9.502	21.65	25.53	30	31.68	27.98	23.63	50.32	64.13	55.23	2.89	33.25	39.4	39.22	3	2.25	3
PI_225869	Gonsenheimer No. 412	Denmark	56.264	9.502	15.49	14.69	16.83	17.68	20.48	20.3	26.99	29.8	31.89	3.23	84.25	79.4	80.22	9	2.5	3
PI_225870	Nantes No. 20	Denmark	56.264	9.502	13.38	14	14.25	10	12.75	11.63	31.87	45.5	35.25	3.23	23	23	21	4	3	3.17
PI_225871	Nantes No. 38	Denmark	56.264	9.502	14.13	14.25	16.37	11	12.13	16.5	33.5	41.25	35	3.23	29	24	33.22	3.5	2.83	3.33
PI_225872	Touchon No. 26	Denmark	56.264	9.502	18	23.63	23	25.75	23.25	21	43.75	62.5	42.25	3.56	50.5	48	38.5	20	3	3
PI_225937		Sweden	60.128	18.644	23.25	20.63	19.12	24.25	24	26.63	52	70.25	52.87	3.23	51	49.5	53	15.5	2.67	3
PI_226043	San Nai No. 1954.8	Japan,_Akita	39.719	140.102	21.25	20	16.75	25.75	23.75	21	45.87	71.5	41.87	1.23	42	44.5	43	5.5	3	3.33
PI_226464	14770	Iran,_Fars	29.104	53.046	21.25	20.63	21.5	28	25.5	25.38	43.99	51.8	40.56	3.56	53.25	58.4	40.22	8	2.83	3
PI_227116	Sweetcrop	New_Zealand	-40.901	174.886	23.63	17.63	21.37	26.5	20.75	18.75	47.75	68.75	46	3.23	62	64	56.5	9	2.33	3.17
PI_230723		Netherlands	52.133	5.291	21.13	20.38	18.12	23.75	20.25	20	46.25	64.75	43.25	2.54	45.5	49	48.5	7	2.33	3.17
PI_234619	Cape Market	South_Africa,_Limpopo	-23.401	29.418	26.49	21.53	17.66	27.68	25.14	26.51	45.99	61.13	43.06	2.89	41.25	51.4	42.22	1	2.33	2.83
PI_234620	Chantenay	South_Africa,_Limpopo	-23.401	29.418	31.5	28.38	30.75	28.25	35.75	35.63	45.8	70.03	51.89	3.23	35.75	38.6	57.78	9	3	2.67
PI_234621	Oxheart	South_Africa,_Limpopo	-23.401	29.418	15.25	15.75	13.37	14.13	10.5	14.46	34.75	34.5	34.37	2.89	31.5	27.5	22	3	3	2.17
PI_234622	Taranaki Improved	New_Zealand	-40.901	174.886	15.13	11.5	16.12	20.25	19.75	16.13	34.5	43	38.62	2.56	23	33	33.5	2	2	3.17
PI_242385	NA	United_States,_Maryland	39.046	-76.641	26.76	25.19	23.8	28	28.48	31.01	37.4	88.52	65.18	3.17	NA	NA	NA	NA	2.75	4
PI_249535	Nantesa	Spain	40.464	-3.749	23.82	34.86	28.5	34.35	37.14	34.46	45.82	66.13	42.23	3.56	55.25	59.4	42.22	5	3	2.83
PI_254552	Zardak Tabur (Zardak = c	Afghanistan,_Kabul	34.555	69.207	25.38	30	29.12	41	40.25	31.25	49.87	79.75	57.5	2.56	52.5	49	53	5.5	2.33	3.17
PI_256065	1	Afghanistan,_Kabul	34.555	69.207	10.99	13.36	13.83	18.68	19.98	12.63	43.32	55.8	35.39	2.89	34.25	34.4	38.22	4	2.67	2
PI_256066	2	Afghanistan,_Kabul	34.555	69.207	13.75	16	17.75	25.5	19	17.5	35.97	100.37	30.56	3.23	15.75	17.6	27.78	0.5	NA	NA
PI_261613	D 74	Spain	40.464	-3.749	14.5	11	12.37	19.5	11.75	16.13	35.5	32	35.25	2.89	36	36	27.5	2.5	2.5	3
PI_261614	St. Valerio	Spain	40.464	-3.749	22.15	24.53	28.33	34.68	28.48	27.8	49.65	92.13	44.73	3.23	63.25	71.4	59.22	6.5	3	3.17
PI_261646	Nakumura Senkofuto	Japan	36.205	138.253	17.88	18.5	18.62	19.75	21	19.5	41.25	47	31.87	2.56	58.5	46	32	3.5	3	2.17
PI_261647	MS Langum	Japan	36.205	138.253	21.75	22.25	23.75	27.75	31.75	29.88	40.5	56.25	40.25	2.56	54	56.5	52.5	11.5	3.5	2.75
PI_261648	Kokubu	Netherlands	52.133	5.291	21.75	20.25	23.37	24.75	26.38	26.63	52.25	80.75	49.5	1.56	56	60.5	74	5	2.67	3.67
PI_261650	High Carotene	Netherlands	52.133	5.291	23.38	28	24.75	28.25	33	29	47.25	72.75	47.12	3.06	47.5	55	46	8	2.33	3
PI_261781	Primerough 1	France,_Ville-de-Paris	48.857	2.352	18	19.13	19.37	29	25.25	20.25	42	50.5	40.5	2.56	32.5	34.5	36.5	3.5	2.67	2.67
PI_261782	Rouge la Merveille	France,_Ville-de-Paris	48.857	2.352	24.13	23.38	27.37	26.75	29.25	27.5	47.75	76.75	47.87	2.89	50.5	52	58	9.5	3.17	3.83
PI_261783	Rouge Muscade	France,_Ville-de-Paris	48.857	2.352	25.02	26.27	26.2	26.25	27.25	24.25	39.4	60.52	52.18	3.23	NA	NA	NA	4	NA	NA
PI_263016	DC 56001	United_Kingdom,_England	52.356	-1.174	29.25	28	28.75	28.75	35	24.5	51.75	71.25	48.62	3.23	47.5	56.5	55	6.5	2.83	3.67
PI_263019	D 267	United_Kingdom,_England	52.356	-1.174	26.5	27	29.87	35.5	31.25	29.75	55.25	95.5	58	2.89	62.5	65	72	16	2.17	2.5
PI_263022	Long Red Stump	United_Kingdom,_England	52.356	-1.174	11.38	12.63	13.37	17.38	17.75	14.5	29.12	27	26.37	3.23	23.5	20.5	29.22	3	1.67	3.17
PI_263023	Kiel Red	United_Kingdom,_England	52.356	-1.174	16.5	16.13	17.12	19.38	19.5	16.5	26.75	36	33.87	3.56	26.5	31	37.5	9.5	2	2.67
PI_263024	Gonsenheim	United_Kingdom,_England	52.356	-1.174	14.38	14.63	13.12	14.25	11.13	10.38	35.5	41.25	35.37	3.23	21	23	19.5	4	2	3.17
PI_264232	Chantenay Red Cored	France	46.228	2.214	23.75	22.63	25.87	29.5	27.63	30.75	55.5	87.25	53.75	3.89	50.5	48.5	60.5	8.5	3.17	2.83
PI_264233	Claudia (Earliest Nantes)	France	46.228	2.214	18.38	18.5	19	23.5	18.5	20.25	34.5	49	34.12	3.73	35.5	41.5	32.5	5	2.33	3
PI_264234	Flakkee	France	46.228	2.214	21.5	18.25	22.5	25.5	25	28.5	43	59	49.25	3.23	40	41	42.5	11	2	2.67
PI_264235	Horn Red Apple	France	46.228	2.214	18.25	18.38	17.37	18.5	24.25	22.75	41.87	59	47.75	3.23	36	37.5	33.5	4.5	1.67	3.17
PI_264236	Nantes Improved A 17	France	46.228	2.214	19.5	18	20	18.13	20.75	25.75	37.75	54.25	45.87	3.89	36	42	37	8.5	2.33	2.67
PI_264237	Vertou	France	46.228	2.214	20.65	16.53	18.16	18.02	20.81	21.63	46.15	62.8	40.73	3.56	33.25	49.4	47.22	3.5	2.17	3.17
PI_264238	Giant Chantenay	France	46.228	2.214	32.13	29	28.12	32	33.75	25.25	61.25	102	59	3.56	49.5	56	54	18.5	3.67	2.17
PI_264543	Kintoki	Japan,_Osaka	34.694	135.502	24.49	24.19	20.5	27	28.5	27	32.59	30.9	29.32	3.23	19.25	20.4	NA	5.5	2	3.75

PI	Name	Origin	est.latitude	est.longitude	early_height_1	early_height_2	early_height_3	early_width_1	early_width_2	early_width_3	late_height_1	late_height_2	late_height_3	disease_score	late_width_1	late_width_2	late_width_3	stand_count	harshness	sweetness
PI_264669	Mohren Bauers Kieler Du	Germany	51.166	10.452	16	16.88	20.25	19.5	15	16.5	33.87	40.75	36.25	2.89	27	24.5	30.5	5	1.83	2.83
PI_267090	Imaki Surk (Vieaki Surkh)	Tajikistan	38.861	71.276	22.5	21.25	24.12	32.75	26.75	28	35.62	51	32.75	3.23	26	29	27	5.5	4.17	3.5
PI_267091	Mirzoi Zholtaya 304	Soviet_Union,_Former	32.808	35	26.25	25.75	26.37	31.5	26.63	29	39.37	49.75	38.87	3.23	45.5	33.5	30.5	8	3.67	3
PI_268382	292	Afghanistan,_Kabul	34.555	69.207	15.5	16.38	17.25	23.25	26.38	23.75	20.97	52.98	30.68	3.23	NA	NA	NA	4.5	2.5	3
PI_269316	Vertou	Sweden	60.128	18.644	12.75	15.13	15.25	15.5	15.5	14.88	30.87	41.75	34.5	3.23	32	36.5	21.5	3	3	3.67
PI_269318	London Torg II	Sweden	60.128	18.644	25.5	25.5	21.5	26.25	38.25	28.75	49	83.5	50.5	3.23	60.5	62.5	69	9	2.83	3.17
PI_269319	Nantes	Sweden	60.128	18.644	17.5	19.75	22.12	20	21.63	22.75	34.25	55	44.62	3.23	37.5	31	37	12	2.67	3
PI_269321	Regulus II	Sweden	60.128	18.644	18	15	17.5	17.5	22.75	19.13	37.75	49.75	36.25	2.89	36.5	38	38	7.5	2.67	3
PI_269322	Amsterdamer	Sweden	60.128	18.644	17.38	15	14.87	13.25	17.75	19.88	29.87	60.25	27.87	3.23	23.5	26.5	23	10	NA	NA
PI_269487	846	Pakistan	30.375	69.345	20.38	16.5	18	23.5	21.88	24.75	43.25	66.75	46.87	3.23	26	37	37	13.5	NA	NA
PI_269488	915	Pakistan,_North-West_Fron	34.953	72.331	25.88	30.75	28.75	35.25	37.5	43.25	49.87	84.5	54.75	3.89	64.5	70.5	65.5	18	3.5	3.75
PI_271044		India	20.594	78.963	35.13	36.25	36.62	36	49.5	45.5	61.25	114.5	66.62	3.89	77.5	77.5	60.5	6.5	NA	NA
PI_271470	Gajer	India,_Gujarat	22.259	71.192	30.63	35.75	31.87	38.25	46.75	42.5	56.97	87.98	67.68	3.23	NA	NA	NA	29.5	2.75	1.25
PI_274789		India,_Delhi	28.704	77.102	25.5	28.75	23.5	34.25	32.25	33.5	51.75	78.5	51.87	3.56	61.5	57.5	60.5	6	3.33	3
PI_276325	Nantes Ndr. Munkegaard	Denmark	56.264	9.502	21.88	19.88	19.75	21.75	21.5	23.38	39.25	49	40.5	3.56	51	54.5	51	7	2.67	3.33
PI_277285	Champion Scarlet Horn	India,_West_Bengal	22.987	87.855	17.88	18.75	20.25	23.5	25.75	21	40.25	49.5	36.5	3.23	51	43	46.5	7.5	3	3.33
PI_277668	Amsterdam Forcing	Netherlands	52.133	5.291	10	11	15	16.5	18.25	21.25	27.5	31.5	29.62	2.56	25	30	24	13.5	3	2.17
PI_277669	Amsterdam Vollegronds	Netherlands	52.133	5.291	16.88	14.63	18.12	19.75	23.5	18	27.25	26.5	27.5	2.23	27.5	25.5	25.5	10.5	2.67	3.33
PI_277709	Amsterdammer Bak	Netherlands,_South_Hollan	52.021	4.494	20.25	18.75	17.37	24.75	27.5	28.5	38.62	46.25	37	2.73	55.5	69.5	57	12.5	3.33	3.17
PI_277711	Springtime	Netherlands,_Limburg	51.443	6.061	17.5	16.75	17.12	19.5	21.25	21.75	31	38	31.5	3.23	44	37.5	41.5	11.5	3	3.25
PI_279776	Balady	Egypt,_Giza	30.013	31.209	33.26	32.69	31.8	32.99	23.84	38.51	NA	NA	NA	3.04	NA	NA	NA	6	NA	NA
PI_279777	2	Egypt,_Giza	30.013	31.209	33.85	24.6	27.2	35	40.5	37.25	48.59	48.9	39.32	3.04	68.25	70.4	60.22	6.5	NA	NA
PI_282480	Osinskaya	Soviet_Union,_Former	32.808	35	28	29.88	24.5	30.5	31.25	35.5	46.25	70.75	46.87	3.23	51.5	59	59	12	3.17	3
PI_284700	London Torg Kampe	Sweden	60.128	18.644	22.88	24.25	22.5	32.5	33.38	30.75	50.37	68.5	49	3.56	51	54	60.5	11	2.67	1.83
PI_284701	Regulus Imperial	Sweden	60.128	18.644	23.38	24.75	20.87	25.38	26.75	27.88	46.62	70	46.25	3.06	47	48	51	11.5	2.33	3.17
PI_285612	Amager	Poland,_Warszawa	52.23	21.012	18.75	19	20.37	25.25	23.75	21.75	46.75	69.75	50.75	2.89	44.5	46.5	43.5	4.5	2.67	3.67
PI_285613	Amsterdamska	Poland,_Warszawa	52.23	21.012	21	22.13	21	26.75	27	26.75	46	62.25	47.12	2.89	35.5	38.5	41.5	8.5	2.83	3
PI_285614	Lenka	Poland,_Warszawa	52.23	21.012	18.69	20.77	14.53	16.32	18.18	18.05	30.47	45.7	34.73	2.23	32.75	45.6	41.78	1.5	2.67	3.5
PI_285615	Londynska	Poland,_Warszawa	52.23	21.012	25.5	25.5	27.75	32.63	35.5	36.25	57.62	87	58.12	3.23	76	79.5	73	16	3	2.67
PI_285616	Nantejska	Poland,_Warszawa	52.23	21.012	24	22.75	17.37	27.5	31	24.25	43	55	43.87	3.56	48	52	48.5	11	2.33	3
PI_285617	Perfekcja	Poland,_Warszawa	52.23	21.012	24.38	28.13	27.12	27.25	33.75	30.75	46.75	69.5	49	3.89	48	62	56	18	2.67	3
PI_285618	Pierwszy Zbior	Poland,_Warszawa	52.23	21.012	20.75	18.13	17.25	33.38	32	27.5	39.5	57	37.87	3.56	44.5	51	42.5	9.5	3	3
PI_285619	Biala Zielonoglowa PZHR	Poland,_Warszawa	52.23	21.012	20.38	27.75	28.87	30.75	29	33.13	53.25	83.25	52	3.56	65.5	66.5	57	9	1.33	2.33
PI_285620	Biala Zielonoglowa SWHI	Poland,_Warszawa	52.23	21.012	28.63	31.75	27.5	36.25	41.5	32.75	54.5	95.5	59.87	3.56	64.5	61	67.5	17.5	1.5	2.25
PI_285621	Lobberychska Busczynski	Poland,_Warszawa	52.23	21.012	27.5	26.75	29	41.5	29.13	25.75	61	97	61.75	2.89	51	57.5	59	24	2.17	2.67
PI_285622	Lobberychska SWHN	Poland,_Warszawa	52.23	21.012	28.38	29.63	30.75	37.25	36.5	32.25	60.87	98.75	60.75	3.23	67	65.5	63	20.5	2.5	3
PI_285623	St. Valery	Poland,_Warszawa	52.23	21.012	25.75	26.38	31.25	35.25	34.75	29.75	57	79.75	55.87	3.23	72	68.5	67.5	18	3.33	3.67
PI_287518	IW 1949	India,_Jammu_and_Kashmi	33.778	76.576	17.24	15.06	16.7	27.5	33.02	34.49	38.04	49.6	26.32	3.33	43.75	46.6	40.78	8	NA	NA
PI_294079	Heian-sanzun-ninjin	Japan,_Kanagawa	35.448	139.642	27.38	28.13	28.5	31.5	34	34.75	51.25	77	51.87	2.73	68	68.5	74.5	9.5	2.67	3.17
PI_294080	Kinko-sanzun-ninjin	Japan,_Kanagawa	35.448	139.642	24.88	28	27.62	29	29.5	31.25	48.37	76.25	50.87	2.73	59.5	66	58.5	9	2.5	3.33
PI_294081	Kurenai-sanzun-ninjin	Japan,_Kanagawa	35.448	139.642	16.75	19.25	17.5	24.25	17.75	20.75	39	57.5	36.12	2.56	42	44	38	3.5	3.67	4
PI_294082	M.S.-sanzun-ninjin	Japan,_Kanagawa	35.448	139.642	22.75	22.5	25.75	33.25	31.88	34.75	41.25	52.75	43.37	3.23	60.5	61.5	62	15	3.83	3.5
PI_294083	Senko-sanzun-ninjin	Japan,_Kanagawa	35.448	139.642	19.15	21.36	24.33	29.35	32.48	24.46	43.15	51.47	39.23	2.89	49.25	46.4	54.22	5.5	3	3.33
PI_294084	Kokubu-senko-onoga-nir	Japan,_Kanagawa	35.448	139.642	24	24.5	21.25	28.38	26.25	28.5	49.62	80.5	52.87	2.89	60.5	65	66.5	14.5	1.83	3
PI_294086	Aichi-gosun-ninjin	Japan,_Kanagawa	35.448	139.642	20	21.25	18.75	25.75	29.75	24.25	40.87	58.5	43	3.06	37.5	40.5	45.5	17	4	3.83
PI_294087	Senko-gosun-ninjin	Japan,_Kanagawa	35.448	139.642	30	28.88	30.25	36.38	33	36.38	46.97	73.37	52.73	3.23	61.75	63.6	61.78	17.5	3.83	3.67
PI_294088	Tokinashi-gosun-ninjin	Japan,_Kanagawa	35.448	139.642	19.75	24	27.12	25	34.13	32.75	50.47	86.7	52.06	2.89	79.75	75.6	80.78	8.5	2.75	2.5
PI_294089	Tokyo-gosun-ninjin	Japan,_Kanagawa	35.448	139.642	20.25	18.75	24.5	27.25	21.25	23.5	43.25	53.5	47.75	3.06	49	42	48.5	8	3.25	3.25
PI_294090	Sapporo-futo-ninjin	Japan,_Kanagawa	35.448	139.642	24	25.03	24.75	23.75	28.25	28.5	49.87	74.75	49.25	2.89	59.5	62	64	17.5	2.33	3.5
PI_295861	Zanahoria silvestre	Spain	40.464	-3.749	-4.06	-1.31	-0.6	4.04	9.47	7.86	23.59	21.9	24.32	2.33	25.25	32.4	20.22	2	NA	NA
PI_295862	7	Spain	40.464	-3.749	9.82	9.19	11.16	15.88	23.81	19.63	20.19	13.25	19.68	3.23	NA	NA	NA	2.5	NA	NA
PI_306588	Takii's Scarlet Chantenay	Japan,_Kyoto	35.012	135.768	18.5	23.13	22.12	28.25	30.63	33.88	52.62	83.75	53.37	3.56	51	59.5	58	14.5	3	3
PI_306810	Topweight	New_Zealand	-40.901	174.886	20.38	27.5	21	23.25	26.75	25.63	50.37	77.25	50.25	3.06	48.5	53.5	53	9	2.5	2.83
PI_319858	Early Scarlet Wonder	Japan,_Kyoto	35.012	135.768	14.13	19.25	18.75	23.5	17.75	17.38	18.4	27.52	41.18	3.56	NA	NA	NA	13	2.5	3
PI_319859	Heian Long Scarlet Wond	Japan,_Kyoto	35.012	135.768	23.5	24	26.25	22.25	25.75	24.75	39.97	54.06	35.5	3.56	36.75	30.6	26.78	10	2.17	3
PI_319860	Takii's Scarlet Wonder	Japan,_Kyoto	35.012	135.768	21.69	24.6	23.37	21.99	24.51	20.71	32.4	69.52	42.68	3.56	NA	NA	NA	17	2.5	4
PI_321688	Kintoki (Early strain)	Japan,_Osaka	34.694	135.502	23	22.13	23.12	25	25.5	33	42.19	43.25	41.18	3.56	NA	NA	NA	14.5	2.5	3
PI_324240	Fortuna	Sweden,_Malmohus	55.99	13.596	20.13	22.88	19	20.13	25.13	28	33.12	41.5	32	3.23	36	37	28.5	16.5	3.67	4
PI_324241	Minerva	Sweden,_Malmohus	55.99	13.596	18.25	17.25	20.12	22.88	20.38	18.13	38.62	48.25	40.5	3.56	41.5	36	37	17.5	2.17	3
PI_325985	K 1582	Russian_Federation,_Krasno	56.015	92.893	22.13	26.5	24	24.25	24.63	25.25	51.5	82.5	50	3.23	64.5	52	62.5	9	3.17	3.17
PI_325987	K 1645	Lithuania	55.169	23.881	21.38	20.63	17.37	23.25	24.38	27	45	61.25	44.12	2.56	44.5	50.5	53	9.5	2.17	3.5
PI_325988	K 1653	Lithuania	55.169	23.881	21.63	21.88	22.5	16	25	25.5	42.12	60.75	39	3.73	45.5	48	52	10.5	3.33	2.83
PI_325989	Nantskaja 4	Russian_Federation,_Mosco	55.756	37.617	23.5	25.25	25.62	24.75	30.75	32.75	44	57.75	46.37	3.56	51.5	62	60.5	12	3.33	3.5
PI_325990	Gribovskaja 514	Russian_Federation,_Mosco	55.756	37.617	25.5	26.63	22.37	24.88	25.5	28.88	48.5	68.25	46.87	3.56	56	53.5	55.5	9.5	3	3
PI_325991	Valeria 5	Russian_Federation,_Mosco	55.756	37.617	21	20.63	23.75	21.75	20	27.13	48	69.75	46.12	2.56	50	46	49	6	2.5	3
PI_325993	Parizskaja (Parisian) Karo	Russian_Federation,_Mosco	55.756	37.617	23.88	18.88	21.12	20.75	23.25	23	47.62	74.75	48	2.56	32.5	38	42.5	10.5	2.75	3
PI_325994	Losinoostrovskaja 13	Russian_Federation,_Mosco	55.756	37.617	24.38	24.5	24.75	19.25	23	22.25	48	67.5	48.25	2.56	50	50.5	49	8.5	2.75	3
PI_325995	Shantene skvirskoe	Ukraine	48.379	31.166	20.5	15.75	21.75	22.5	19.75	19.75	39.62	60.5	35.75	3.23	34	37.5	27	5	3.17	3.5
PI_325996	Nesravnennaja	Ukraine	48.379	31.166	22.63	25	27.75	24	24.38	26.38	48.87	72	49.75	2.89	47.5	53	54	8	3.25	4
PI_325997	K 1706	Ukraine	48.379	31.166	14.13	14.88	15.25	13.75	15.25	13.38	36.12	50.25	37.37	2.23	36.5	34	24.5	2	2.67	3.33
PI_325999	Nantskaja Harkovskaja	Ukraine	48.379	31.166	15.63	19.63	19.75	25.25	20.75	21	37.75	45.5	37.75	3.23	49.5	51.5	47	12.5	3.33	3.67
PI_326000	Nantskaja 14	Ukraine	48.379	31.166	16.25	19.13	22.62	19.25	18.75	20.5	38	54.75	38.75	2.89	38.5	41	44	9	3	3.17
PI_326001	Gavrilovskaja	Ukraine	48.379	31.166	24.88	28.38	24.37	25.75	28.13	32.75	53.62	91.5	50	3.23	54.5	53.5	55	12.5	3.75	4
PI_326002	Biriucekutskaja 415	Ukraine	48.379	31.166	25.25	25	23	24.75	24	22.5	49.5	67.5	50.62	3.89	55	51	56	16	3.5	3.5
PI_326003	Shantene 2461	Russian_Federation,_Altay	50.618	86.22	27.5	30	27	29	30	33.75	48.47	89.7	58.23	3.56	48.75	54.6	57.78	13	2.67	3.33
PI_326004	Altajskaja ukorocennaja	Soviet_Union,_Former	32.808	35	29.13	28.88	23	23.88	27.5	29	55.12	71	50.37	3.39	59.5	78	65	9	3.5	2.25
PI_326005	Hibinskaja	Russian_Federation,_Murm	68.959	33.083	22.25	27.25	27.62	24.25	30.25	30.38	46.25	61	46.75	3.39	54	64.5	61	16.5	3.83	3.67
PI_326006	Geranda	Russian_Federation,_Voron	51.675	39.209	34.63	29.88	30.75	31.25	34	31.13	57	85	55.25	3.23	67.5	73.5	69	18.5	3.33	3.5
PI_326007	Leningradskaja	Russian_Federation,_Lening	59.934	30.335	33.5	31.25	27.5	30.5	29.5	31.88	54	70.25	49.87	3.56	51	57.5	55	29	3	2.67
PI_326009	Mirzoi Krasnaia (red) 228	Uzbekistan	41.377	64.585	17.88	17.25	18.75	20.25	23	28.5	36	42	40.62	3.23	48.5	36.5	42	11	2.83	3
PI_326010	Mshaki-surk	Tajikistan	38.861	71.276	25.5	22	23.62	34.38	29.63	31.25	42.25	54.5	45.75	3.56	34.5	41.5	46	5.5	3	3.33
PI_326011	Tushon	Lithuania	55.169	23.881	23.25	25.5	23.5	29.13	25.25	26.25	41.25	56.25	45.75	3.56	51	49	43.5	12.5	3.25	3.25
PI_326012	Nantskaja Goriskaja	Georgia	41.715	44.827	18.25	21.13	17.5	21.5	20.5	23.25	42.37	62.75	41.37	3.23	36.5	36.5	30	4.5	2.5	3.5
PI_326013	Sibirskaja Krasnaja	Russian_Federation,_Omsk	54.988	73.324	27	27.75	26.5	33.5	35.13	32.25	62.87	106	57.62	3.23	62.5	75	62	16.5	2.5	3
PI_326014	Leninakanskaja	Armenia	40.069	45.038	18.75	23.13	21	22.75	21.25	22.5	46.37	60.75	42.12	2.89	28.5	35.5	41	3.5	3.5	3
PI_339252	8Hv-2	Turkey,_Eskisehir	39.767	30.526	17.75	17.38	17.37	23.24	18.11	29.05	39.5	64.75	37	3.56	38	36	24	5	2.67	2.5
PI_341204	Flakkee	France	46.228	2.214	20.63	15.38	19.75	19.5	26.25	26.96	44	69.75	44.12	2.89	40	37.5	42	7	1.5	3.25
PI_341205	Nantaise B	France	46.228	2.214	16.88	15.5	13.37	18	24.13	21.5	35.37	40.75	29.5	3.23	44	31	33	5.5	2.17	3.83
PI_341206	Nantaise de Maininet	France	46.228	2.214	20.88	19.88	17.75	16.63	22.75	25.25	32.75	42.5	33.87	3.23	33.5	42	33	6.5	3	3.75
PI_341207	Parisienne Forcer	France	46.228	2.214	18.38	16.38	16.37	18.75	20.63	22.63	37.25	39.5	35.87	3.23	38	36	30.5	7	3	3
PI_341208	Prim Rouge	France	46.228	2.214	25.25	25.13	26.5	35	34.75	29.75	49.37	61.75	46.25	3.06	62.5	65	48	11	3.83	3.83
PI_344072	22681	Turkey,_Gaziantep	37.066	37.378	32.35	30.27	38.2	27.99	35.84	34.05	64.64	87.03	57.23	2.89	27.75	38.6	29.78	4	2.5	3
PI_344110	Londynska	Poland	51.919	19.145	23.75	18.5	17.87	24.75	30.75	23.75	46.75	70.25	48.12	2.89	34	40	59	8.5	4	3.5
PI_344360	Yerli Havuc	Turkey,_Trabzon	41.003	39.717	21.88	19.25	19.12	19.5	28.75	26.5	45.25	61.5	43	2.89	59.5	47.5	54.5	5.5	3.33	4
PI_344361	Renklin	Turkey,_Konya	37.875	32.493	29.88	28.38	26.5	21.75	29.13	27.63	59.12	94.75	50.87	3.23	47	55	55.5	15.5	2.67	3.33
PI_344362	Kirmiza	Turkey,_Icel	36.812	34.641	24.5	21.38	20.5	16.25	21.75	19.75	46.8	104.7	45.06	2.96	13.75	11.6	6.78	11	1.5	3
PI_357975	Stipski	Macedonia	41.609	21.745	20.75	27.13	19.75	17.75	26.88	27.63	42.87	59.25	48.12	3.56	38	37	29.5	13	3	2.5
PI_357979	Domasen	Macedonia	41.609	21.745	23.25	24.38	20.75	21.13	25.38	21	46.62	63	49.37	3.23	44.5	51.5	41	10.5	2	3.5
PI_357980	Kumanovski	Macedonia	41.609	21.745	20	19.75	22.75	22.75	22.75	25.25	34.12	56.75	43.12	3.23	31	49	44.5	6	4.33	3.67
PI_357981	Prilepski	Macedonia	41.609	21.745	22.88	22.5	18.12	20.75	24.5	22.5	33.75	41.75	39.12	3.73	40.5	41.5	47	8	3.5	3.5
PI_357982	Domasen	Macedonia	41.609	21.745	19.75	17.5	16.12	18.25	18.88	20.96	36.5	43	35.25	3.56	30	35	46.5	9.5	4	3.5
PI_357983	Stara sorta	Macedonia	41.609	21.745	18.13	14.25	14.62	27.75	19.25	23.75	35.75	71.25	36	3.23	34	28	28.5	0.5	3.75	3.75
PI_357984	Mesten	Macedonia	41.609	21.745	14.15	18.19	14.83	18.35	16.14	14.8	32.65	37.13	29.06	2.89	16.25	12.4	11.22	0.5	3.83	2.83
PI_357986	Tap	Macedonia	41.609	21.745	18.82	16.86	15.83	15.68	14.64	19.63	27.5	54.5	30	2.89	13	17.5	16	1.5	3.5	3.25
PI_357987	Dolg	Macedonia	41.609	21.745	16.5	15.75	17.25	18	22.5	18.13	40.25	53.25	38.25	2.56	35	39	31	2.5	4	3.67
PI_357988	Bitolski	Macedonia	41.609	21.745	26.88	23.13	20.87	28.5	28.75	38.63	45.15	42.8	38.06	2.56	57.25	53.4	59.22	8.5	3.83	3.67
PI_368620	Dolg	Macedonia	41.609	21.745	21.88	15.63	18.12	22.75	18.5	20.63	38.19	63.75	47.18	3.23	NA	NA	NA	2	2	3.33
PI_368622	Obicen	Macedonia	41.609	21.745	16	12.88	14	24	29.13	25.3	39.5	39	36	3.23	41.5	43.5	44.5	8.5	3.5	3.5
PI_368623	Vratnick	Macedonia	41.609	21.745	13.63	15	12.25	25.38	20.75	17.88	27.75	30.75	25.75	3.23	30.5	42	51.5	5	2.5	3
PI_369349	Sapporo Large Long	Japan,_Kyoto	35.012	135.768	17.88	14.25	12	18.38	16.5	16.96	36.75	47.25	37.5	2.23	35	36.5	40	1	2.67	3
PI_370505	Mesten	Macedonia	41.609	21.745	22.13	19.13	23	27.75	30.25	34.5	44.5	73.75	43	3.23	59.5	54.5	60	15.5	3.25	3.75
PI_378882	Konfrix	Germany	51.166	10.452	13	11.63	13.37	16.38	17.38	18.25	30.75	35.75	29	3.23	20.5	25	23	6.5	3.33	3.67
PI_379325	Trgoviski	Serbia	44.017	21.006	24	24.5	22.75	26.75	24.38	28.75	40.87	58.75	33.25	3.89	47.5	48	55.5	16.5	3.17	3.67
PI_379327	Siljasti	Serbia	44.017	21.006	25.88	27.13	32.5	32.13	35.5	25.13	52.5	63.5	49.25	3.56	65	71.5	49.5	16	2	3
PI_379328	Prizrenski	Serbia	44.017	21.006	27.25	22.13	25.37	35.25	33.63	37.75	44.72	83.56	50	3.23	30.75	16.6	30.78	16.5	1.83	2
PI_379329	NA	Macedonia	41.609	21.745	24.26	23.44	18.3	28.5	31.48	38.51	39.19	54.25	35.68	3.67	NA	NA	NA	NA	2.5	2
PI_418967	Sian, Chi-Tou	China,_Shaanxi	35.394	109.188	25.25	24	21.5	24.13	24.13	23.5	44	60.5	40.5	3.23	40.5	47	37.5	14	2.33	2.83
PI_419109	Huang pi, hu lo pu (yellov	China	35.862	104.195	24.63	26.38	26.5	33.5	35.88	35.25	43	72.75	44	2.89	38.5	63.5	67.5	13	NA	NA
PI_419184	Pan Te Hung	China	35.862	104.195	19.38	22.5	19.12	23	24.25	20.5	34.25	51.25	39.75	2.89	45	39	35	5	2.67	4
PI_430524	VIR 205	Azerbaijan	40.143	47.577	16.25	18.75	19	22	19.5	21.88	42.62	47.5	36.75	3.56	26.5	37.5	33.5	5.5	NA	NA
PI_430525	VIR 233	Afghanistan	33.939	67.71	13.74	10.56	8.2	25.5	27.52	28.99	30.31	40.75	29.32	3.83	35.5	29	34	5.5	NA	NA
PI_430527	Murzon	Uzbekistan	41.377	64.585	21.5	24.5	21.5	26.5	29.38	28.13	38.5	64.25	40.25	3.23	43	65.5	36.5	8	3	2.83
PI_430529	Mirzon Zeltaja	Uzbekistan	41.377	64.585	26.75	26	21	26.5	31	31.38	33.37	34.25	33.5	3.56	46	40.5	50	8	3.75	2.75
PI_430530	Msakisupx	Tajikistan	38.861	71.276	27.75	25.25	19	28.75	28.38	27.3	46.5	67.25	38.12	3.73	61.5	47.5	56	4.5	3.75	3.5
PI_430532	VIR 2207	Russian_Federation,_Dages	42.143	47.095	25.13	18.13	18	21.75	27.13	25.88	46.14	72.03	48.89	2.89	62.75	63.6	42.78	4.5	NA	NA
PI_430533	VIR 2263	Iraq	33.223	43.679	18.5	19.75	19.37	29.75	21.75	24.13	38.75	55.5	42.5	3.39	37.5	29.5	45	16.5	2.33	2.67
PI_430534	VIR 2278	Afghanistan	33.939	67.71	16.88	15.38	17.12	20.25	24.88	23.5	34.5	46	37	3.23	41.5	30.5	37	9	3	3
PI_432899	Chang hong	China	35.862	104.195	18.25	22.13	17.75	22.75	21	32.3	42.25	61	37.75	3.23	36.5	41	38.5	6	3.17	3.67
PI_432900	Yellow carrot 11	China	35.862	104.195	20.88	22.25	18.75	19.5	21.25	21	37.5	51.5	34.5	3.56	29	32.5	38.5	4.5	2.75	3
PI_432901	Yellow carrot 12	China	35.862	104.195	27.25	27.38	21.5	27	28	29.13	46	62	45.25	3.56	40.5	50	38	8	3	2.17
PI_432902	Xiao fing	China	35.862	104.195	23.63	20.88	24.87	23.38	27	32.38	48.25	77.5	49.87	3.56	43	42.5	35.5	14.5	3.67	2.17
PI_432904	H 001	China	35.862	104.195	22.63	25.75	21.62	25.25	30.5	30.25	49.75	67	47.25	3.23	41.5	49	54.5	23	2.17	3.17
PI_432905	Sa 102	China	35.862	104.195	17.5	19.38	20.25	14.63	21	19.75	47.5	56.75	37.5	3.23	42	41.5	52.5	8	2.5	2.83
PI_432906	Sa 103	China	35.862	104.195	29.13	31.5	30.12	30	30.38	28.5	46.5	63	50.5	3.23	74.5	63	59	28.5	2.75	3
PI_451752	Lange witte groen kop	Netherlands	52.133	5.291	26.5	27.63	25.37	19.25	32.25	30.63	48	71.75	49.5	3.56	51.5	58	63.5	15	2.17	3.17
PI_451755	Lange gele stomper	Netherlands	52.133	5.291	16.63	17.63	16.5	22.5	18.13	17	41	56.25	41.5	2.89	30.5	28.5	28	5.5	NA	NA

PI	Name	Origin	est.latitude	est.longitude	early_height_1	early_height_2	early_height_3	early_width_1	early_width_2	early_width_3	late_height_1	late_height_2	late_height_3	disease_score	late_width_1	late_width_2	late_width_3	stand_count	harshness	sweetness
PI_451756	Lange gele Koe	Netherlands	52.133	5.291	22.63	24.5	14.12	28	33	28.25	32.99	58.13	35.89	2.23	35.25	33.4	43.22	8	3	2
PI_451757	Flakee Samo	Netherlands	52.133	5.291	14.75	20.13	21.25	31.25	33.5	28.75	39.47	55.37	43.56	3.39	57.75	56.6	55.78	7	2.75	4
PI_451759	Opbrenger	Netherlands	52.133	5.291	18.38	18.13	18	25	23.25	26.63	44.25	53.5	40.5	3.73	44.5	40.5	45	7	1.75	3.25
PI_451761	Mommersteeg Lange gel	Netherlands	52.133	5.291	11.5	17.5	20.5	19.5	13	20.88	36.75	45.25	34.75	3.56	33	29	30.5	11	NA	NA
PI_458858	Ca 2-2	Russian_Federation,_Mosco	55.756	37.617	22.38	23.13	25	24.5	27.5	26.75	47.12	62.75	42.75	3.39	64.5	60	53	14.5	3.17	3.5
PI_458859	Ca 3-1	Russian_Federation,_Mosco	55.756	37.617	21.88	19.25	16.62	19.75	21.75	20.13	35.25	52.5	32.75	2.89	47	32.5	34.5	3.5	3	3.33
PI_458860	Natez	Russian_Federation,_Mosco	55.756	37.617	29.13	30.38	26.87	26.5	31.25	31.25	50.75	66.5	47	3.39	43	66	69.5	16	2.67	3.67
PI_478370	O 70	China,_Xinjiang	42.525	87.54	26	23.75	22.12	33.25	34.5	30.75	42.5	61	35.25	2.56	49.5	50.5	49	7	2.5	3.17
PI_483349	Spring F1 (Spring Favor)	Korea,_South,_Seoul	37.567	126.978	26.5	25.75	26.75	30.25	29.75	27.5	48.25	87	52.5	2.73	59	58	56	9	3	3
PI_483352	Summer-5	Japan	36.205	138.253	30	26.88	25.5	34.25	31.63	34.5	53.25	78.75	55.75	2.23	64.5	59.5	63	8	2.5	3.17
PI_502914	Long red blunt	Germany	51.166	10.452	20.38	19.13	20.25	24.25	22.63	21.25	39.25	58	39.75	2.89	53.5	43	52.5	4.5	2.83	3
PI_502918	Rondo	Germany	51.166	10.452	18.69	16.27	18.2	16.65	18.68	20.38	36.14	31.7	29.39	2.23	30.75	34.6	38.78	5.5	2.5	2.83
PI_502919	Rotin	Germany	51.166	10.452	21.5	22.13	18.37	24.75	23.13	25.63	33	46.75	43.5	3.56	37.5	39.5	35.5	6	4	3.5
PI_502920	Short Early Duwicker	Germany	51.166	10.452	16.38	19.25	18.37	21	20.88	15.88	39.25	48.25	40.37	2.89	53	53.5	48.5	5	2.83	3.17
PI_503342	Nantskaja Jygeva	Estonia	58.595	25.014	20.63	26.88	25.62	21.38	30	27	42.37	44	39	3.23	44	44	43	8	2.75	3
PI_503343	Garduoles	Lithuania	55.169	23.881	24.13	23.88	26.62	27	27.25	31.88	45	59.25	42.37	3.23	52	60.5	59	11.5	2.17	3
PI_506444	80821	Kazakhstan	48.02	66.924	28.5	38.88	32.25	35.75	37.5	42.75	65.87	101.5	60.62	3.89	68	71	82	30	3	2.75
PI_506445		Kazakhstan	48.02	66.924	19.25	25.5	20	20.32	32.18	24.55	56.3	95.37	47.89	2.89	35.75	86.6	66.78	1	2.75	3.5
PI_508470	Spring favor	Korea,_South	35.908	127.767	28	23	30.12	26.38	29.25	31.75	53.87	83.75	50	3.23	77	78	66.5	15	2.5	2.83
PI_508471	Summer favor	Korea,_South	35.908	127.767	22.75	31.38	28	29.5	27	36.25	52.25	90.25	54.87	2.89	67.5	45	47.5	6.5	3.17	3.67
PI_508472	Prolific 5	Korea,_South	35.908	127.767	22.75	20	21.5	28.5	27.5	30.25	44.5	62.25	45.37	2.56	18	37.5	35	3	2.83	3.17
PI_508473	Red core	Korea,_South	35.908	127.767	23.75	26.75	23.62	41.5	32	31.5	44	67.25	44.5	2.89	54.5	57.5	52.5	4.5	2.17	3.17
PI_509434	Kirmizi havuc (Red carrot	Turkey,_Sivas	39.751	37.015	25.75	25.38	21.37	25.13	24.75	27.75	42.75	52	51	3.56	42.5	33	52	7	NA	NA
PI_509435	Havuc (carrot)	Turkey,_Mardin	37.313	40.734	21.38	17.25	17	25.38	26.81	19.8	34.3	21.56	26	2.96	34.75	28.6	15.78	1.5	NA	NA
PI_515990	VII-1-158	Hungary	47.162	19.503	16.25	17.25	21.5	17.75	16.88	23.25	42.75	68	45.25	3.23	44	35.5	45	11	2.17	3.17
PI_515993	VII-1-235	Hungary	47.162	19.503	23.13	24.63	20.5	25.25	19.13	23.5	47	56.75	51	2.73	53	48	40.5	7.5	1.83	3.17
PI_515994	VII-1-239	Hungary	47.162	19.503	29.63	29.25	24.37	26.5	22.5	28.5	47.14	43.56	42.5	3.23	49.75	57.6	62.78	29.5	2.25	2.75
PI_515997	VII-1-251	Hungary	47.162	19.503	20	17.75	20.87	24.75	25	20.13	54.69	40.52	45.68	3.23	NA	NA	NA	10	1.75	3.75
PI_515998	VII-1-252	Hungary	47.162	19.503	31.38	30.25	30.87	30.5	32.5	29.13	51.87	80	57	3.56	68	76.5	62	19	2.5	3.17
PI_515999	VII-1-253	Hungary	47.162	19.503	28	28.25	31.62	25	23.75	28.75	46	70.25	48.12	3.23	64	61	59.5	26	1.67	3
PI_516000	Keszthelyi Hengeres	Hungary	47.162	19.503	24.5	25.88	22.5	26	24.25	29.63	41.25	47.75	34.75	3.23	48.5	45.5	48	18.5	3.25	3.5
PI_516001	Fertodi Voros	Hungary	47.162	19.503	28	24.13	24	18.5	25.5	26.5	45.62	63.5	45	3.56	49	52	55.5	14	2.33	3.17
PI_522173	Benifuku Fuyugosi 5 Sun	Japan,_Hokkaido	43.22	142.863	21.38	23.75	22.62	26	27.25	27.75	52.75	76	47.75	2.23	46.5	43	54	9.5	2.83	3.33
PI_522174	Benifuku 4 Sun	Japan	36.205	138.253	21.63	21.13	22	22.75	21.38	21.25	42.75	60.75	47	2.73	30	38.5	32	7	3.5	3
PI_522175	Benifuku 625	Japan	36.205	138.253	26.25	21.88	26.75	27.75	29	27.88	45.75	67	48.12	2.73	50	48	39	12.5	2.83	3.17
PI_522176	Benifuku Harumaki 5 Sur	Japan	36.205	138.253	22.38	26.63	26.5	31.25	31.63	32.63	53.25	77.75	55.12	2.73	56.5	81	69.5	11.5	3.67	4
PI_522177	Kyokujitu 5 Sun	Japan	36.205	138.253	31.75	25.88	28.25	29.75	40.75	34.75	59.5	85.25	58.5	2.56	87.5	84	74	14.5	3.33	3.5
PI_531324	Bacsbokodi	Hungary	47.162	19.503	17.38	16	11.5	15.88	17.48	20.3	37.25	48	32.75	3.23	23.5	29	22.5	2.5	2.5	2.75
PI_531325	Fertodi Voros	Hungary	47.162	19.503	20.25	17.5	21.12	24	20	17.88	31.5	48	34.5	3.23	23	32	27.5	7.5	2	3.5
PI_531326	Kiskunhalasi	Hungary	47.162	19.503	18.63	20.88	20.37	28.25	26	24.25	42.12	49.25	42.75	3.23	29	46	53	7	3	3
PI_535882	Jawa	Poland	51.919	19.145	21.88	25.63	23.37	24	27.38	28.5	44	74.25	45.12	3.73	40	46	46	15	2.17	3.5
PI_535884	Lenka	Poland	51.919	19.145	24.5	24.38	23.25	23.88	27.75	23.5	43.62	67.75	44	3.39	56	59.5	56.5	19.5	3.5	3.67
PI_535886	Pierswzy Zbior	Poland	51.919	19.145	21.38	22.25	23.5	23.25	27.75	23.13	46	64.75	42.5	3.56	56	52	41.5	11	2.5	2.5
PI_540418	U044	Uzbekistan	41.377	64.585	11.74	14.06	13.2	19.5	11.02	16.49	26.31	25.25	25.82	2.33	27.5	30.5	25	8.5	NA	NA
PI_540419	Mirzoe Mushtak	Uzbekistan	41.377	64.585	32	30.63	30	37.75	38.25	37.25	46.25	80	40.25	3.73	63	53.5	59	36.5	3.83	3.33
PI_540422	U110	Uzbekistan	41.377	64.585	13.5	17	15	25	21.81	18.8	37.32	43.8	32.89	2.89	NA	NA	NA	2	2.5	2.5
PI_632381	Yellow Belgian	United_States,_California	36.778	-119.418	25.75	24.38	28	35	39.14	34.96	48.25	87.75	49.62	3.23	35.5	49	34.5	11	3	2.75
PI_632382	Burpees Oxhart	United_States,_Pennsylvani	41.203	-77.195	25.5	24.25	20.25	24.75	21.75	27.3	50	79.75	52.75	3.56	47	47	47	16.5	2.83	3.33
PI_632385	Imperator Long Type Sho	United_States,_California	36.778	-119.418	24	25	22.75	29	29.88	22.25	41.75	63	45.87	3.73	46	58	43.5	8.5	2.67	3.17
PI_632386	James Intermediate	United_States,_California	36.778	-119.418	23.38	23.63	17.37	27.5	26.75	27.13	49.5	67	48.62	3.89	47.5	45.5	43.5	13	3	3
PI_632387	Tablequeen	United_States,_Illinois	40.633	-89.399	23.63	23.38	26.12	26.5	21.5	37.75	41.25	54.25	42	3.89	51	53	50	17.5	2.5	2.67
PI_632389	Dutch Horn	Netherlands,_North_Hollan	52.521	4.788	23.5	23.38	21	21.75	22.25	25.13	38.12	46.75	44.37	3.23	31.5	34.5	44.5	9	3.5	3.75
PI_632390	Long Imperator II	United_States,_Michigan	44.315	-85.602	28	28	23.12	28	30.75	33.38	42.25	56	37.62	3.73	38.5	60	54.5	23	3.5	1.75
PI_632391	Long Imperator 58	United_States,_California	36.778	-119.418	25.5	26.75	20.87	21.63	25.75	23.63	41.5	64.75	47.75	3.89	53.5	48	49	17	4	3.5
PI_632393	Waltham Hicolor	United_States,_California	36.778	-119.418	26	29.25	26.87	36.5	30.88	32.25	48.37	80	51.12	3.23	65.5	60.5	61.5	18.5	4	3.83
PI_632394	Popsicle	United_States,_Pennsylvani	41.203	-77.195	26.5	25.75	26.75	26.88	34.5	32.13	50.25	61.75	49.5	2.56	50	45	52	12	3.5	3.83
PI_632395	Red Cone	United_States,_California	36.778	-119.418	20.25	20	18.37	25.75	18.75	19.5	32	37.25	36.25	2.89	40	29	27.5	7	2.83	2.83
PI_634650	Chantenay/Model	United_States,_Pennsylvani	41.203	-77.195	25.13	25.13	27.75	22.75	28.38	26.63	53.87	72.75	52.25	2.23	54.5	42	56	7	2.5	3.33
PI_634651	Chantenay Long Type	United_States,_Minnesota	46.73	-94.686	28.25	28.63	30.75	30.25	27.38	30	55.75	86	56.12	3.73	58.5	56	57	36	2.67	3
PI_634652	Long Orange	United_States,_Missouri	37.964	-91.832	28.25	31.88	28	26.63	33.25	32.38	53.75	81.75	53.12	3.54	59	60.5	57	12.5	2.17	3
PI_634653	Tendersweet	United_States,_Missouri	37.964	-91.832	29.5	28.38	28.75	23.13	25.75	29.38	55.62	75.25	54.37	3.23	56	52	48	20.5	2.25	2.5
PI_634655	Woods Scarlet Intermedi	United_States,_Virginia	37.432	-78.657	25.13	18.88	21.62	31	30.14	32.46	46.5	53.75	40.12	3.56	36	34	35	4.5	2.83	4
PI_634656	Airliner	United_States,_Michigan	44.315	-85.602	26.88	27.13	26.87	28.75	26.88	27.5	50.87	74.25	49.75	3.73	51	46	45	20.5	3	3.5
PI_634657	Nantes Tip Top	Netherlands,_North_Hollan	52.521	4.788	25.63	26.38	27.25	28	25.63	28.5	45.62	48.25	38.12	3.56	45	38	37.5	22	3.67	2.83
PI_634658	C Saint Fiacre	France	46.228	2.214	22.5	20.88	21	19.25	20.13	24.13	39	48.75	37.25	2.89	38.5	40	31.5	5.5	2.75	2.75
PI_642755	French Forcing	United_States,_California	36.778	-119.418	17	13	13.5	18.88	24.48	19.8	34.37	44.75	31.75	3.23	31	27	32.5	6.5	2.5	1.75
PI_642756	Amsterdam Coreless	Netherlands,_North_Hollan	52.521	4.788	14.88	14.38	11.87	19.25	23.63	20	27	42.75	19	3.23	27	31	34.5	8	3.5	2.25
PI_642757	Early Golden Ball	Netherlands,_North_Hollan	52.521	4.788	25.13	21.63	24.75	23.63	26.13	24.88	45.75	66	42.75	3.23	53.5	40	47	11	2.5	2.17
PI_642759	Best of All	United_Kingdom	55.378	-3.436	27.38	25.5	25.75	26.5	35.63	26.38	51.12	85.25	45.87	2.89	45	48	45.5	12	2.75	2.5
PI_642760	Wonderkugel	Switzerland	46.818	8.228	20.5	22.63	20.87	27.38	28	29.63	43.75	61.25	41	3.56	55	53	50.5	20.5	2.5	3.25
PI_642761	Oxheart	United_States,_California	36.778	-119.418	25.25	29.13	29.25	38.25	34.13	28.25	50.37	78.25	50.62	3.23	53	64	51	16.5	2.83	3.17
PI_643114	White Belgian	United_States,_California	36.778	-119.418	31.88	31.63	29.87	33.25	31.63	31.63	57.37	87.5	57.75	3.89	65	72	75	18.5	NA	NA
PI_643115	Tiny Sweet	United_States,_Minnesota	46.73	-94.686	24.5	27.5	29.37	31.5	28	28.75	41.75	61.25	42.5	3.56	55	44.5	45.5	18.5	4	3.75
PI_643116	Chanticleer	United_States,_Connecticut	41.603	-73.088	34	34.75	32.25	25.13	32.88	38.63	55.37	80	55.12	3.89	75.5	63	66	24	2	3.5
PI_643117	Fidler's Exhibition	Unknown	41.626	-79.674	30.25	31.13	29.37	28.25	31.75	36.75	59	87.5	55.75	3.56	69	77.5	63.5	14	2.5	3
PI_643118	Selected Long Orange Im	United_States,_New_York	40.713	-74.006	21.5	20.25	21.12	20.63	25.88	26.25	47.62	68.75	42.25	3.89	66.5	57	54	11.5	2.5	2.75
PI_643119	Tilques	France	46.228	2.214	24.5	27.38	27.87	25	31	22.13	50.12	61	52.5	3.89	48.5	54.5	54.5	12.5	2.5	2
PI_643120	Prinant	United_States,_New_York	40.713	-74.006	19.75	15.75	14.12	15.38	15.81	24.63	34.75	71.5	26.75	3.89	25.5	32.5	32.5	5.5	3	2.5
PI_652118	Nantaise (A Forcer)	France	46.228	2.214	22.25	24	18.12	21.5	28	23.38	38.5	50	39.12	3.56	48.5	34	47.5	18	3.75	3.5
PI_652119	Crimson Wonder	Japan	36.205	138.253	21.49	22.53	27.16	26.13	31.5	31	39.78	52.94	39.32	3.56	24.25	18.4	9.22	6.5	1.5	2
PI_652121	Shin Kurodane Gosun	Japan,_Ibaraki	36.342	140.447	33.25	29	23.75	38.25	31.5	36.38	58	90.25	54.5	2.23	61.5	59.5	71	8	3	3.25
PI_652122	Sone	Japan,_Ibaraki	36.342	140.447	26.75	31.25	28	32.13	34	32.25	58.5	89.5	61.5	1.89	68	57	60	9.5	2.17	3.33
PI_652123	US Harumaki Gosun	Japan,_Ibaraki	36.342	140.447	26.38	27	22.87	26.25	26.75	22.25	49.5	65.75	52.25	3.06	46.5	43	52	10	3.17	2.83
PI_652124	Tamahata Yonsun	Japan,_Ibaraki	36.342	140.447	26.5	26.13	28.37	28.5	25.75	35.13	47.5	63.75	40	3.23	48	47	52	8.5	3.83	3.17
PI_652125	Sankimaki Sanzun	Japan,_Ibaraki	36.342	140.447	27.75	28.38	28.75	27.25	24.75	29.25	44.5	71	46.25	2.73	48.5	48	52.5	13	3.75	3.5
PI_652127	Tokinashi Gosun	Japan,_Ibaraki	36.342	140.447	23.5	28	25.37	29.38	33.75	29.25	46	72.25	49.12	3.06	61	54.5	54	10.5	3.33	2.67
PI_652128	Nakamura Senkou Futo	Japan,_Ibaraki	36.342	140.447	27.88	31.25	27.87	33.75	38	28.13	55.25	88.25	53	3.23	73	65.5	74.5	14.5	2	3.25
PI_652129	Manpukuji Senkou Oona	Japan,_Ibaraki	36.342	140.447	41.25	41.88	38.5	41.5	41.5	42.63	68.75	108.5	69.37	2.54	80.5	93	85	30.5	2.67	3.33
PI_652130	Senkou Sapporo Futo	Japan,_Ibaraki	36.342	140.447	27.38	30.63	34.37	33.75	29.38	31.75	55.25	89.75	60	2.89	73.5	72.5	60.5	18.5	2.75	3
PI_652131	Sapporo Futo	Japan,_Ibaraki	36.342	140.447	24.25	27.38	26.25	22.63	31.25	26.13	48.25	65.25	50	1.89	61	69	46.5	9	2.5	3.67
PI_652132	Sapporo Futo	Japan,_Ibaraki	36.342	140.447	29.88	26.38	28.62	28	29.75	26.88	54.5	79.5	47.25	2.56	65	76	68	16	2.33	3.17
PI_652135	Shinshuu Senkou Oonaga	Japan,_Ibaraki	36.342	140.447	25.13	27.88	27.62	29.5	30.5	23.5	57.25	89.25	52.37	2.23	62	65	51	7	2.83	3.33
PI_652136	Shin Kuroda Gosun	Japan,_Ibaraki	36.342	140.447	30.63	30	38.5	35.75	39.5	36.75	64	96.5	57	2.89	75.5	71	68	23.5	2.5	3
PI_652138	Yoshino	Japan,_Ibaraki	36.342	140.447	23.51	20.44	25.55	25.32	21.51	22.71	35.4	25.52	39.68	4.56	NA	NA	NA	7	1.5	3
PI_652147	Rosal	Netherlands	52.133	5.291	27.75	27.5	23.25	21.75	23.5	26.13	50	70.25	43.62	3.56	56	67.5	62	21.5	3	2.75
PI_652148	Vitaminnaja 6	Russian_Federation	61.524	105.319	17.75	16	19.25	14.5	20.38	19	35.12	45.25	38.25	3.23	33.5	46.5	38	10.5	3	3.33
PI_652149	Olympia	Czech_Republic,_Central_Bo	49.878	14.936	17	14.13	17.62	17.25	19.14	18.96	34.87	41.5	32.12	3.23	19.5	20	27.5	4	3.67	3
PI_652152	Yates Market King	United_Kingdom	55.378	-3.436	16.15	14.53	12.5	18.18	16.98	12.8	23	63.75	27.37	2.56	24	31.5	23.5	4.5	NA	NA
PI_652155	Flakker	Hungary,_Pest	47.448	19.462	16.35	15.77	14.7	18.49	16.84	21.51	33.3	44.37	34.73	2.23	10.75	40.6	36.78	2.5	2.83	2.67
PI_652157	Vesta Vennaja	Soviet_Union,_Former	32.808	35	21	23.88	21.37	29.5	18.25	18.25	42.87	62.25	42.5	3.23	43	46	36	9.5	3	3
PI_652158	Landrace 1982:404	Georgia	41.715	44.827	24.63	26.38	24.12	25	23.25	27.88	44.25	64	50.87	3.56	46.5	45	49	16	2.83	3
PI_652160	Amsterdam Grace	Denmark	56.264	9.502	19.5	17.5	20.75	22.75	22	21.5	39.75	43.5	33.75	2.89	51	38.5	39.5	15.5	2.5	3
PI_652163	Vita Longa	Netherlands	52.133	5.291	21.75	21.13	22.62	14.25	19.5	20.63	47.87	77	42.75	2.56	44.5	39	33	7	4	2.5
PI_652164	Regina	Denmark	56.264	9.502	15.38	20.5	15.5	13.13	15.75	20	34.37	79.25	38.12	2.89	29	27	38.5	7	2.67	3
PI_652165	Superpak	Netherlands	52.133	5.291	14.75	14.88	16.12	16.75	16.5	18.75	30.5	32.25	30.62	2.89	30.5	18.5	26.5	8	2.17	3.17
PI_652166	Superno	Netherlands	52.133	5.291	18.5	22.25	20	21.25	23	21.88	36.75	55.75	44	2.23	37	50.5	47.5	12	2.83	2.17
PI_652167	Formula	Netherlands	52.133	5.291	21.75	24.38	20.87	18	25.5	28.75	44.87	77	55.25	2.89	50	42.5	57.5	33	3	3.5
PI_652169	Decca	France	46.228	2.214	16	16.5	18.49	17.75	10.75	15.26	30.25	30.5	27.77	2.56	30	19	18.5	4	3.17	3.5
PI_652170	Tantal	France	46.228	2.214	20.25	22.13	21	22.5	22.25	18	36.37	54.75	27.87	3.56	35	35.5	30.5	12	2.5	3
PI_652171	Karotan	Netherlands	52.133	5.291	17.38	16.38	19	18	23.38	15.5	34	39.25	34.87	2.89	38.5	50.5	35.5	3.5	2	3
PI_652173	Amsterdamer Finger	United_Kingdom,_England	52.356	-1.174	11.25	10.34	9.65	11	11.45	14.55	22.87	31.5	23.87	2.04	21.5	14.5	12	5	2.5	3.5
PI_652174	Voros Orias	Hungary,_Pest	47.448	19.462	19.5	21.75	24.25	22	24.75	25	46.75	67	45.37	3.56	46	57.5	56	11	2	3
PI_652175	Slendero	Netherlands	52.133	5.291	20.88	22	23.25	21.75	22.5	20.75	36.75	41.5	37.75	3.23	36.5	38.5	42	10.5	2.67	3.17
PI_652177	Regulus II	Netherlands	52.133	5.291	18.13	16.25	19.66	19.75	23	21.5	27.5	41.25	36.5	2.89	31.5	43	40	2.5	2.33	2.83
PI_652179	Danvers 126	United_States	37.09	-95.713	26.38	24.88	30.25	26.25	25.5	26.38	45.37	68.25	48.75	3.23	46.5	56.5	46.5	5	2	2.75
PI_652180	Moskovskaja Zimnjaja	Soviet_Union,_Former	32.808	35	22.82	19.53	23.16	21.68	29.14	21.3	44.32	69.8	46.73	3.23	32.25	53.4	34.22	3	2.33	3.33
PI_652188	Ping Ding	China,_Beijing	39.904	116.407	27.75	25	26.37	23.63	27.38	31.13	51.5	72.75	47.25	3.56	53	58.5	60	15.5	2.5	2.75
PI_652200	Shantene	Kazakhstan,_Alma-Ata	43.222	76.851	27.75	29.38	30.37	35.75	40.25	34.5	48.25	71.5	52.25	3.73	59	60.5	62	29	2.5	2.17
PI_652201	Ames 19034	Kazakhstan,_Alma-Ata	43.222	76.851	31.63	31.63	29.12	33.75	34.5	38.5	53.62	89	52.75	3.56	61	66	66	36.5	2.67	3
PI_652202	Artek	Moldova	47.412	28.37	32.25	30.25	26.75	37.75	30	34.5	48.25	72.25	53.75	3.56	48	46.5	48	14	2.5	2.83
PI_652203	Biriucekutskaja	Russian_Federation	61.524	105.319	31.88	29.38	28.5	23.38	37.5	36	50.75	83.5	54	3.56	42	54	63.5	13	2.5	3
PI_652204	Konservnaja 63	Moldova	47.412	28.37	32.75	30.25	27.87	32.5	28.75	37	51.75	90	51.87	3.56	52.5	58.5	67.5	18	3	3.17
PI_652205	Nantskaja Gorijskaja	Georgia	41.715	44.827	26.5	24.88	26.25	28.75	26.13	26.88	47.5	64.75	47.12	3.56	45	53	52	15	2.17	2.83
PI_652206	B92-76	Bulgaria	42.734	25.486	31.13	28	28.75	28.5	27	31.25	47.25	68	47	3.56	45.5	49.5	52.5	12	3	3.25
PI_652207	Bian gan hong	China	35.862	104.195	20.75	22.13	22	25	23.5	28.38	46	73	44.62	2.23	43.5	58	50.5	7	2	2.5
PI_652209	A ke su hu luo bu	China	35.862	104.195	25.63	29	27.5	20.75	35	32.25	53.37	76.25	52	2.89	57	52	47	13	2.5	2.75
PI_652210	Tu lu fan hu luo bu	China	35.862	104.195	32	30.88	25.75	30.25	28	38.13	57.12	70.25	50.5	3.23	43.5	45.5	52	18.5	2.17	2
PI_652211	Ha mi huang pi hu luo bu	China	35.862	104.195	34.25	33.88	31.12	34.13	35.63	37.5	51.12	85.25	50.75	3.89	63.5	56.5	54.5	16.5	2.5	2.5
PI_652212	Ha shi hong pi hu luo bu	China	35.862	104.195	26.5	24.75	21.87	25	24.13	26.75	52.5	70	48.12	3.23	41.5	39.5	43.5	7.5	2.25	2.5
PI_652217	Nantes forto	Nepal	28.395	84.124	27.25	28.5	23.12	30.5	31.75	27.25	47.25	73.75	46.25	3.89	44	42	38	15.5	2.75	2.75
PI_652231	Nantaskaya 4	Armenia	40.069	45.038	29.63	25.25	23.25	23.5	23.25	25.75	51.12	69.25	46.75	3.56	47.5	43.5	32.5	4.5	3.5	2
PI_652232	Leninakanian-6	Armenia	40.069	45.038	24	22.13	28.25	22.88	24.25	31.96	54.75	76.25	52.5	3.56	54	52	37.5	10.5	4	2.17
PI_652243	IIHR 089	Turkey	38.964	35.243	40.5	39.13	41.5	39.65	38.84	43.38	53.99	69.8	56.56	3.56	76.25	77.4	78.22	21.5	2.5	2.25
PI_652244	IIHR 091	Turkey	38.964	35.243	33.82	25.86	29.33	45	26.25	29.3	44.49	49.71	37.5	3.89	50.25	53.4	52.22	3	2.75	2.5
PI_652245	IIHR 161	India	20.594	78.963	35.26	36.69	34.55	36.32	39.18	35.05	NA	NA	NA	3.23	NA	NA	NA	13.5	NA	NA
PI_652246	IIHR 162	Russian_Federation	61.524	105.319	41.38	37	33.5	33.25	41.5	39.75	61.59	144.9	64.32	3.23	74.25	61.4	69.22	25.5	1.83	2.33

PI	Name	Origin	est.latitude	est.longitude	early_height_1	early_height_2	early_height_3	early_width_1	early_width_2	early_width_3	late_height_1	late_height_2	late_height_3	disease_score	late_width_1	late_width_2	late_width_3	stand_count	harshness	sweetness
PI_652247	IIHR 163	Russian_Federation	61.524	105.319	32.38	33.63	31.5	38.65	42.18	37.05	NA	NA	NA	3.89	NA	NA	NA	16.5	2	2.33
PI_652249	IIHR 165	Russian_Federation	61.524	105.319	31.88	30.75	30	24.32	31.84	25.71	48.78	63.44	47	3.56	53.25	51.4	54.22	30.5	2.17	2.83
PI_652254	IIHR 193	India,_Uttar_Pradesh	26.847	80.946	32.13	29.75	31	30.25	35.38	32	61.14	95.37	57.23	3.04	55.75	62.6	63.78	11	2.25	3.25
PI_652255	IIHR 195	India,_Uttar_Pradesh	26.847	80.946	34.75	35.5	37	35.38	40.25	36	55.14	72.03	50.23	3.89	74.75	79.6	66.78	44	1.67	3
PI_652258	Pusa Yamadagni	India,_Delhi	28.704	77.102	27.5	21.25	26	26	20.48	17.63	35.65	35.47	32.89	3.23	44.25	41.4	41.22	4.5	2.25	2.5
PI_652268	POL176408	Poland,_Bielsko	49.822	19.058	21.38	23.25	23	30.65	34.51	31.05	30.4	44.52	29.68	3.56	NA	NA	NA	22	2	3
PI_652269	POL176409	Poland,_Poznan	52.406	16.925	27.38	28.25	27.62	33.5	34	32.75	42.47	51.06	38	3.89	79.75	67.6	46.78	17.5	1.5	3.33
PI_652277	VIR 1609	Mongolia	46.862	103.847	22.88	21.75	19.62	14.25	19.81	19.13	30.25	61.25	37.5	2.89	32	28.5	22.5	2.5	2.83	2.17
PI_652278	VIR 1713	Kyrgyzstan	41.204	74.766	13.51	13.19	13.55	18	16.73	15.76	15.44	0.75	18.93	2.56	NA	NA	NA	0.5	NA	NA
PI_652279	VIR 1769	Russian_Federation	61.524	105.319	18.88	16.13	10.62	14.13	11.25	19.3	24.5	29.5	24.25	3.23	23	25.5	19	5	2.5	2.83
PI_652280	VIR 1772	Russian_Federation	61.524	105.319	24.25	24.13	17.25	23.5	17.25	26.88	50	68.25	48.75	3.56	42	51	47.5	8	3.17	3.17
PI_652281	VIR 1826	Russian_Federation	61.524	105.319	27.13	30.25	26.75	29.5	27	25.25	45.37	60.25	48.5	2.89	49	58	54	8	2.17	2.83
PI_652282	VIR 1843	Albania	41.153	20.168	29	26.63	24	27.38	23.25	26	47	68.5	45.5	3.23	57	64	46.5	10.5	2.33	3.17
PI_652283	VIR 1847	China	35.862	104.195	29.5	23.88	23.5	23.88	20.63	21.96	44.25	64.25	46.37	3.23	51	42	41.5	9.5	2.33	3.5
PI_652284	VIR 1851	Bulgaria	42.734	25.486	18.75	20.13	23.5	19.63	23.98	25.3	39.12	53.75	40.25	3.23	42.5	33	34.5	4.5	1.67	2.33
PI_652286	VIR 2052	Bulgaria	42.734	25.486	17.13	17.75	14.12	17.5	17.75	15.5	41.49	56.47	37.73	3.56	49.25	41.4	30.22	2	2.17	2.67
PI_652287	VIR 2080	Armenia	40.069	45.038	28.25	26.75	25.5	27.75	27.25	31.25	53.75	63.75	47.25	3.73	44	47.5	56.5	9.5	2.33	3
PI_652288	VIR 2086	Kazakhstan	48.02	66.924	15.75	13.88	15	16.13	18.14	16.96	32	36	27.5	3.23	15.5	17	13.5	1	3.17	2.17
PI_652335	S107	Syria	34.802	38.997	22.38	22.5	22.75	24.75	22.75	26.75	39	43.5	39	3.23	26	21	20.5	6	2	2
PI_652400	T120	Turkey,_Denizli	37.783	29.096	27.25	30.75	29.75	31.75	27.75	32.63	52	80	50.25	3.89	42	44	43.5	13.5	NA	NA
PI_652401	T121	Turkey,_Denizli	37.783	29.096	30.25	32.5	32.62	27.25	31.38	29.63	57	80.25	53	3.89	43	48	69.5	47.5	NA	NA
PI_652402	T123	Turkey,_Denizli	37.783	29.096	35.75	40.63	36	31.5	30.5	27	54	85.25	54.25	3.89	64.5	77	71	98.5	NA	NA
PI_652403	T124	Turkey,_Denizli	37.783	29.096	35.13	35	30.62	32.63	33	30.5	57.25	91.5	59.62	3.89	66	59.5	65.5	49	NA	NA
PI_652404	T125	Turkey,_Denizli	37.783	29.096	23.88	18.5	22.62	19.63	19.13	17.5	44.5	54	42.5	3.23	44	54	24.5	9.5	3.83	3.33
PI_652405	Isparta	Turkey,_Denizli	37.783	29.096	24.38	24.38	28.12	27.75	22.75	20.38	49.25	75	52	3.56	43.5	44	50	10.5	NA	NA