frontiers

# Characterization of the USDA *Cucurbita pepo*, *C. moschata*, and *C. maxima Germplasm* Collections

**Christopher O. Hernandez [1], Joanne Labate [2], Bob Jarret [3], Kathleen Reitsma [4], Jack Fabrizio [5], Kan Bao [6], Zhangjun Fei [6,7], Rebecca Grumet [8] and Michael Mazourek[5*]**

[1] *Department of Agriculture Nutrition and Food Systems, University of New Hampshire, Durham, NH, 03824 USA*
[2] *Plant Genetic Resource Conservation Unit, United States Department of Agricultural Research Service, Geneva, NY, 14456 USA*
[3] *Plant Genetic Resource Conservation Unit, United States Department of Agricultural Research Service, Griffin, GA, 30223 USA*
[4] *North Central Regional Plant Introduction Station, Iowa State University, Ames, IA, 50014 USA*
[5] *Plant Breeding and Genetics, Cornell University, Ithaca, NY, 14853 USA*
[6] *Boyce Thompson Institute, Cornell University, Ithaca, NY, 14853 USA*
[7] *U.S. Department of Agriculture-Agriculture Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, 14853, USA*
[8] *Department of Horticulture, Michigan State University, East Lansing, MI, 48824 USA*

Correspondence*:
Michael Mazourek
mm284@cornell.edu

## ABSTRACT

The *Cucurbita* genus is home to a number of economically and culturally important species. We present the analysis of genotype data generated through genotyping-by-sequencing of the USDA germplasm collections of *Cucurbita pepo*, *C. moschata*, and *C. maxima*. These collections include a mixture of wild, landrace, and cultivated specimens from all over the world. Roughly 1,500 - 32,000 high-quality single nucleotide polymorphisms (SNPs) were called in each of the collections, which ranged in size from 314 to 829 accessions. Genomic analyses were conducted to characterize the diversity in each of the species and revealed extensive structure corresponding to a combination of geographical origin and morphotype/market class. Genome-wide associate study (GWAS) was conducted for each data set using both historical and contemporary data, and signals were detected for several traits. These data represent the largest collection of sequenced Cucurbita and can be used to direct the maintenance of genetic diversity and the development of breeding resources, and to help prioritize whole-genome re-sequencing for further GWAS and other genomics studies aimed at understanding the phenotypic and genetic diversity present in *Cucurbita*.

**Keywords: Germplasm, Genotyping-by-sequencing, GWAS, Diversity, *Cucurbita***

## 1 INTRODUCTION

The *Cucurbitaceae* (Cucurbit) family is home to a number of vining species mostly cultivated for their fruits. This diverse and economically important family includes cucumber (*Cucumis sativus*), melon (*Cucumis melo*), watermelon (*Citrullus lanatus*), and squash (*Cucurbita ssp.*)(Ferriol and Picó, 2008). Like other cucurbits, squash exhibit diversity in growth habit, fruit morphology, metabolite content and disease resistance, and have a nuanced domestication story (Chomicki et al., 2020; Paris and Brown, 2005). The genomes of *Cucurbita ssp.* are small (roughly 400 Mb), but result from complex interactions between ancient genomes brought together through an allopolyploidization event (Sun et al., 2017). These factors make squash an excellent model for understanding the biology of genomes, fruit development, and domestication. Within *Cucurbita*, five species are recognized as domesticated. Three of these are broadly cultivated: *C. maxima*, *C. moschata*, and *C. pepo* (Ferriol and Picó, 2008). Few genomic resources have been available for these species; although, draft genomes and annotations, along with web-based tools and other genomics data are emerging (Yu et al., 2022). Already, these resources have been used to elucidate the genetics of fruit quality, growth habit, disease resistance, and to increase the efficiency of cucurbit improvement (Montero-Pau et al., 2017; Zhong et al., 2017; Kaźmińska et al., 2018; Wu et al., 2019a; Xanthopoulou et al., 2019; Hernandez et al., 2020); however, there has yet to be a comprehensive survey of the genetic diversity in the large diverse *Cucurbita* germplasm panels maintained by the USDA within the National Plant Germplasm System.

Germplasm collections play a vital role in maintaining and preserving genetic variation. These collections can be mined by breeders for valuable alleles and can also be used by geneticists and biologists for mapping studies (McCouch et al., 2020). Like many other orphan and specialty crops,there has been little effort put into developing community genetic resources for squash and other cucurbits. The Cucurbit Coordinated Agricultural Project (CucCAP project) was established to help close the knowledge gap in cucurbits. This collaborative project aims to provide genomics resources and tools that can aid in both applied breeding and basic research. The genetic and phenotypic diversity present in the USDA watermelon, melon, and cucumber collections has already been explored as part of the CucCap project, partially through the sequencing of USDA germplasm collections and development of core collections for whole-genome sequencing (Wang et al., 2021, 2018; Wu et al., 2019b). The diverse specimens of the USDA squash collections have yet to be well characterized at the genetic level;although, an elaborate system has been established for classifying squash based on species and various other characteristics.

The classification system used in squash is complex. Squash from each species can be classed as winter or summer squash depending on whether the fruit is consumed at an immature or mature stage, the latter is a winter squash (Loy, 2004). Squash are considered ornamental if they are used for decoration, and some irregularly shaped, inedible ornamental squash are called gourds; however, gourds include members of *Cucurbita* as well as some species from *Lagenaria*—not all gourds are squash (Paris, 2015). Many squash are known as pumpkins; the pumpkin designation is a culture dependent colloquialism that can refer to jack O' lantern types, squash used for desserts or, in some Latin American countries, to eating squash from *C. moschata* known locally as Calabaza (Ferriol and Picó, 2008). Cultivars deemed as pumpkins can be found in all widely cultivated squash species. Unlike the previous groupings, morphotypes/market classes are defined within species. For example, a Zucchini is reliably a member of *C. pepo* and a Buttercups are from *C. maxima*. Adding to the complexity of their classification, the *Cucurbita* species are believed to have arisen from independent domestication events and the relationships between cultivated and wild species remains poorly understood (Kates et al., 2017).

*C. pepo* is the most economically important of the *Cucurbita* species and is split into two different subspecies: *C. pepo* subsp. *pepo* and *C. pepo* subsp. *ovifera* (Xanthopoulou et al., 2019). Evidence points to Mexico as the center of origin for *pepo* and southwest/central United States as the origin of *ovifera*. The progenitor of *ovifera* is considered by some to be subsp. *ovifera* var. *texana*, whereas subsp. *fraterna* is a candidate progenitor for *pepo* (Kates et al., 2017). Europe played a crucial role as a secondary center of diversification for subsp. *pepo*, but not subsp. (Lust and Paris, 2016). Important morphoptypes of *pepo* include Zucchini, Spaghetti squash, Cocozelle, Vegetable Marrow, and some ornamental pumpkins. *C. pepo* subsp. *ovifera* includes summer squash from the Crookneck, Scallop, and Straightneck group, and winter squash such as Delicata and Acorn (Paris et al., 2012).

The origin of *C. moschata* is more uncertain than *C. pepo*; it is unclear whether *C. moschata* has a South or North American origin (Chomicki et al., 2020). Where and when domestication occurred for this species is also unknown; however, it is known that *C. moschata* had an India-Myanmar secondary center of origin where the species was further diversified (Sun et al., 2017). *C. moschata* plays an important role in squash breeding as it is cross-fertile to various degrees with *C. pepo* and *C. maxima*, and can thus be used as a bridge to move genes across species (Sun et al., 2017). Popular market classes of *C. moschata* include cheese types like dickenson, which is widely used for canned pumpkin products, Butternut (Neck) types, Japonica, and tropical pumpkins known as Calabaza (Ferriol and Picó, 2008).

*C. maxima* contains many popular winter squash including Buttercup/Kabocha types, Kuri, Hubbard, and Banana squash (Ferriol and Picó, 2008). This species also sports the world's largest fruit, the giant pumpkin whose fruit are grown for competition and can reach well over 1000 Kg (Savage et al., 2015). Although this species exhibits a wide range of phenotypic diversity in terms of fruit characteristics, it appears to be the least genetically diverse of the three species described (Kates et al., 2017). *C. maxima* is believed to have a South American origin, and was likely domesticated near Peru, with a secondary center of domestication in Japan and China (Nee, 1990; Sun et al., 2017).

In this study, we set out to characterize the genetic diversity present in the USDA *Cucurbita* germplasm collections for *C. pepo*, *C. moschata*, and *C. maxima*. We present genotyping-by-sequencing (GBS) data from each of these collections, population genomics analysis, results from genome-wide association study (GWAS) using historical and contemporary phenotypes, and suggest a core panel for re-sequencing.

## 2 MATERIALS AND METHODS

### 2.1 Plant Materials and Genotyping

All available germplasm were requested from USDA cooperators for *C. maxima* (534 accessions from Geneva, NY), *C. moschata* (314 accessions from Griffin, GA), and *C. pepo* (829 accessions from Ames, IA) respectively. Seeds were planted in 50-cell trays and two 3/4 inch punches of tissue (approximately 80-150 mg) was sampled from the first true leaf of each seedling. DNA was extracted using Omega Mag-Bind Plant DNA DS kits (M1130, Omega Bio-Tek, Norcross, GA) and quantified using Quant-iT PicoGreen dsDNA Kit (Invitrogen, Carlsbad, CA). Purified DNA was shipped to Cornell's Genomic Diversity Facility for GBS library preparation using protocols optimized for each species. Libraries were sequenced at either 96, 192, or 384-plex on the HiSeq 2500 (Illumina Inc., USA) with single-end mode and a read length of 101 bp.

## 2.2 Variant Calling and Filtering

SNP calling was conducted using the TASSEL-GBS V5 pipeline (Glaubitz et al., 2014). Tags produced by this pipeline were aligned using the default settings of the BWA aligner (Li and Durbin, 2009). Raw variants were filtered using BCFtools (Danecek et al., 2021). Settings for filtering SNPs were as follows, minor allele frequency (MAF) $\geq 0.05$, missingness $\leq 0.4$, and biallelic. Three outlier genotypes were found in an initial principal components analysis (PCA) of the *C. maxima* data and were removed, as they were likely not *C. maxima*. Variants were further filtered for specific uses as described below.

## 2.3 Population Genomics Analysis

ADMIXTURE (Alexander and Lange, 2011), which uses a model-based approach to infer ancestral populations ($k$) and admixture proportions in a given sample, was used to explore population structure in each dataset. ADMIXTURE does not model linkage disequilibrium (LD); thus, marker sets were further filtered to obtain SNPs in approximate linkage equilibrium using the "–indep-pairwise" option in PLINK (Purcell et al., 2007) with $r^2$ set to 0.1, a window size of 50 SNPs, and a 10 SNP step size . All samples labeled as cultivars or breeding material were removed from the data prior to running ADMIXTURE. Cross-validation was used to determine the best value for each species. Briefly, ADMIXTURE was run with different values (1-20) and the cross-validation error was reported for each $k$. The $k$ value with minimal cross-validation error was chosen for each species Figure 1. Ancestral populations were then assigned to cultivars using the program's projection feature.

Principal components analysis (PCA) was used as a model-free way of determining population structure. PCA was conducted using SNPRelate (Zheng et al., 2012) on the same LD-pruned data used by ADMIXTURE.

Linkage disequilibrium was calculated in each germplasm panel using VCFtools (Danecek et al., 2011) with the settings "–geno-r2 –ld-window 1000".

## 2.4 Analysis of Phenotypic Data

Historical data were obtained from the USDA Germplasm Resources Information Network (GRIN; www.ars-grin.gov) for *C. maxima*, *C. pepo*, and *C. moschata*. All duplicated entries were removed for qualitative traits, where categories are mutually exclusive, leaving only samples with unique entries for analysis. Two traits: adult and nymph squash bug damage in *C. pepo* where transformed using the boxcox procedure. Contemporary phenotypic data were collected from a subset of the *C. pepo* collection grown in the summer of 2018 in Ithaca, NY. Field-grown plants were phenotyped for vining bush habit at three different stages during the growing seasons to confirm bush, semi-bush or vining growth habit. Plants that had a bush habit early in the season but started to vine at the end of the season were considered semi-bush.

## 2.5 GWAS

Variant data were filtered to MAF and missingness , and then imputed prior to association analysis. LinkImpute (Money et al., 2015), as implemented by the TASSEL (Bradbury et al., 2007) "LDKNNiImputatioHetV2Plugin" plugin was used for imputation with default settings. Any data still missing after this process were mean imputed. The GENESIS (Gogarten et al., 2019) R package, which can model both binary and continuous traits. All models included the first two PCs of the marker matrix as fixed effects and modeled genotype effect ($u$) as a random effect distributed according to the kinship

(**K**) matrix ($u \sim N(0, \sigma_u^2 \mathbf{K})$). Binary traits were modeled using a logistic regression with GENESIS. The kinship matrix was calculated using A.mat from rrBLUP (Endelman, 2011) with mean imputation.

### 2.6 Genomic Heritability

An estimate of genomic heritability ($h_G^2$) (de los Campos et al., 2015) was calculated for all ordinal and quantitative traits using an equivalent model to what was used for GWAS, but without fixed effects. Variance components from the random genetic effect ($\sigma_u^2$) and error ($\sigma_e^2$) were then used to calculate the heritability as $h_G^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$.

### 2.7 Syntenty of Bu putative region in C. pepo and C. maxima

All tools used in the analysis can be found on the Cucurbit Genomics website (http://cucurbitgenomics.org/). A candidate gene for dwarfism, Bu, in C. maxima was elucidated by a previous study (Zhang et al., 2015) and was named Cma_004516. The Cucurbit Genomics Database gene ID of Cma_004516 was identified by using the BLAST tool to align primer sequences used for RT-QPCR in the previous study against the *C. maxima* reference genome. The synteny analysis was done by using the Synteny Viewer tool and evaluating *C. maxima*'s chromosome 3 with *C. pepo*'s chromosome 10 and searching for an ortholog to the candidate gene. The physical position of the *C. pepo* ortholog was identified by searching the gene using the Search tool. All tools used in the analysis can be found on the Cucurbit Genomics Database at cucurbitgenomics.org.

### 2.8 Identification of a Core Collection

Subsets representative of each panel's genetic diversity were identified through running GenoCore (Jeong et al., 2017) using the filtered SNP sets. The GenoCore settings were "-cv 99 -d 0.001".

## 3 RESULTS

### 3.1 Genotyping

Each *Cucurbita* ssp. collection was genotyped using the GBS approach. The collections comprised 534 accessions for *C. maxima*, 314 for *C. moschata*, and 829 for *C. pepo*. Figure 2 shows the geographical distribution of accessions broken down by species. *C. maxima* and *C. moschata* constitute the majority of accessions collected from Central and South America, whereas *C. pepo* accessions are more prevalent in North America and Europe. *C. pepo* had the highest number of raw SNPs (88,437) followed by *C. moschata* (72,025) and *C. maxima* (56,598). After filtering, *C. pepo* and *C. moschata* had a similar number of SNPs, around 30,000, whereas *C. maxima* had an order of magnitude fewer filtered SNPs (1599). This discrepancy may be an artifact of using PstI, a rarer base-cutter previously optimized for GBS of *C. maxima* (Zhang et al., 2015) , rather than ApeKI which was used for *C. pepo* and *C. moschata*. The number and distribution of SNPs across each chromosomes is shown in Table 1. Maps of SNP distribution for each species are shown in Figure 3

### 3.2 Population Structure and Genetic Diversity

Filtered SNPs were used for population structure analysis. Available geographical, phenotypic, and other metadata were retrieved from GRIN and were used to help interpret structure results. Results from model-based admixture analysis are shown in Figure 4 panel a. These data support 10 ancestral groups

(K=10) in C. pepo, 6 in *C. moschata*, and 6 in *C. maxima* in each of the species respectively. Population structure was driven mostly by geography, except in *C. pepo* where the presence of different subspecies was responsible for some of the structure. Commonalities among structure groups are described in Table 2. The first two principal components (PCs) of the marker data are shown in Figure 4 panel b. As with the model-based analysis, PCA showed geography as a main driver of population structure with accessions being derived from Africa, the Arab States, Asia, Europe, North America, and South/Latin America. PC1 in *C. pepo* separates *C. pepo* subsp. *ovifera*, which have a North American Origin, from subsp. *pepo*.

Ancestry proportions from admixture analysis were projected onto cultivars/market types identified in the accessions, which were excluded from the initial analysis used to infer ancestral groups. Cultivars were grouped according to known market class within species to help identify patterns in ancestry among and between market classes. Key market types identified in accessions from *C. pepo* including Acorn, Scallop, Crook, Pumpkin (Jack O' Lantern), Zucchini, Marrow, Gem, and Spaghetti; Neck, Cheese, Japonica, and Calabaza in *C. moschata*; and Buttercup, Kobocha, Hubbard, and Show (Giant squash) in *C. maxima*. These groupings are shown in Figure 5. In general, members of each market class exhibit similar ancestry proportions. In *C. pepo*, market classes from the two different subspecies had distinct ancestry patterns. For example, Acorn, Scallop and Crook market classes are all from subsp. *ovifera* and all of these classes had similar ancestry proportions with roughly 20% of ancestry from the wild *ovifera*. In contrast, market classes within *pepo* had a small percentage of ancestry from wild *ovifera* and more ancestry in common with European and Asian accessions. With *C. moschata*, Neck, Cheese, and Calabaza market classes showed very similar ancestry patterns, whereas the Japonica class was more distinct. Relative to the *C. pepo* and *C. moschata*, the *C. maxima* cultivars were less differentiated from one another.

Results from linkage-disequilibrium analysis are shown in Figure 6.

## 3.3 Analysis of Phenotypic Data

All available historical data from GRIN were compiled. Only traits with 100 entries were considered for further analysis. Filtering resulted in 26 traits for *C. pepo*, 5 for *C. moschata* and 16 for *C. maxima*. Traits spanned fruit and agronomic-related characteristics, as well as pest resistances. The number of records for a given trait ranged from 108 to 822, with an average of 270. Fruit traits included fruit width, length, surface color and texture, and flesh color and thickness. Agronomic data included plant vigor and vining habit, and several phenotypes related to maturity. Pest-related traits included susceptibility to cucumber beetle and squash bug in *C. pepo* and Watermelon mosaic virus (WMV) and powdery mildew (PM) in *C. maxima*. Figure 7 shows the distribution of numeric traits.

## 3.4 Genome-wide Association

Genome-wide association study was conducted for all traits using standard mixed-model analysis. A weak signal was detected in *C. moschata* on chromosome 3 for fruit length. Weak signals were detected in *C. maxima* for fruit ribbing on chromosome 17 and green fruit on chromosome 20. Five phenotypes were significantly associated with SNPs in *C. pepo*: bush/vine plant architecture on chromosome 10 using contemporary and historic data, fruit flesh thickness on chromosome 2, green fruit on chromosomes 2 and 19, and a non-significant, but clear signal for flesh color on chromosome 5. The bush/vine phenotype exhibited the strongest signal. Manhattan plots for the GWAS results are shown in Figure 8, and the corresponding quantile-quantile plots are shown in Figure 9.

### 3.5 Genomic Heritability

### 3.6 Syntenty of Bu putative region in *C. pepo* and C. *maxima*

A candidate gene for dwarfism found in the species *C. maxima* was named Cma_00451623 and corresponds to the gene ID CmaCh03G013600 in the Cucurbit Genomics Database. The gene Cp4.1LG10g05740 on chromosome 10 in *C. pepo* was found to be orthologous to CmaCh03G013600 and coincides with the region significantly associated with the bush/vine plant architecture phenotype identified by GWAS in the *C. pepo* collection.

### 3.7 Development of a Core Collection

A core set of accessions that covered over 99% of total genetic diversity was identified in each of the panels. Roughly 5%-10% of the accessions were required to capture the genetic diversity in the panels (see Figure 10. This amounted to 117 accessions in *C. pepo*, 72 in *C. moschata*, and 72 in *C. maxima*.

## 4 DISCUSSION

Cucurbita *pepo*, *C. moschata*, and *C. maxima*, exhibit a wide range of phenotypic diversity. This diversity was evident in the GRIN phenotypic records for these species. We have demonstrated that there is also a wide range of genetic diversity through genotyping-by-sequencing and genetic analysis of available specimens from the germplasm collections. Thousands to tens of thousands of whole-genome markers where discovered for each species. Clustering of samples and admixture analysis produced results that align closely with known secondary centers of origin in all species. This was especially clear in our analysis of the *C. pepo* collection. *Cucurbita pepo* has its origin in the new world, with a secondary center of diversification in Europe. This pattern was conspicuous in the our PCA analysis. These markers and our analysis of available germplasm have a number of uses for breeding and future experiments aimed at biological insight.

Our GWAS analysis using contemporary and historic plant habit phenotypic data led to the mapping of a locus on chromosome 10 associated with the bush/vine phenotype. This locus is likely the bush gene (Bu) locus that has been finely mapped to this location in previous *C. pepo* studies (Xiang et al., 2018; Ding et al., 2021). Although our GWAS hit does not a constitute a novel gene association, it does demonstrate that the Bu locus, previously mapped in biparental populations, is also the primary driver of the bush phenotype in diverse germplasm. Thus, this locus can likely be applied across a wide array of germplasm. We also demonstrated that this locus is syntenic with the bush gene previously mapped in *C. maxima* (Zhang et al., 2015).

Our data provides many genome-wide markers which could be used to develop marker panels for use in breeding applications, as has been done in other crops (). Possible breeding applications would include marker assisted selection, marker assisted backcrossing, and purity assessment of seedstock using a low density panel; whereas, a medium density panel could be developed for routine genomic selection. Our clustering of samples based on marker data suggest geography is a key driver for overall population structure. When projecting ancestry proportions onto cultivars of known market classes, the ancestry proportions were relatively similar within market class grouping. Although there is genetic diversity within each species, this diversity is constrained within market classes. This suggests that crosses between these market classes would greatly increase the amount of genetic diversity to be leveraged in breeding efforts. Crossing between market classes would come at the cost of bringing in undesirable characteristics with

regard to achieving a specific morphotype associated market class. This cost could be mitigated through the use of markers to recover morphotype expeditiously during pre-breeding.

Genomic selection (GS) was proposed over twenty years ago (), and has since become a standard breeding technique. Yet, to our knowledge, GS is not being used in applied breeding programs working with cucurbits. Studies specifically looking at GS in squash have demonstrated, as is the case in other crop, that GS is a viable breeding method; although, the specific implementation may vary for each program and must take into account the nature of the trait being predicted (). Since cucurbit crops are more space-limited than seed-limited, a predict-part-test-part or sparse testing strategy is potentially an even more efficient strategy in cucurbits than it has been shown to be in grain crops (). Selective phenotyping of resource-intensive quality traits based on marker data to enable prediction is also low-hanging fruit. Our work lowers the barrier to entry for GS in squash, as it provides a set of markers that can be filtered and rapidly converted into an amplicon-based assay for use in target germplasm. This set can then be used for routine genotyping, which is a necessary first step towards implementing GS ().

Our data provides a useful starting point for association studies. In the case where traits are common in the panel, the panel can be phenotyped for a trait of interest and combined with marker data and insight provided by our study. We demonstrated this approach in our association analysis of the bush gene. In the case of a rare phenotype, such as a resistance gene, subsets of the germplasm and markers should be used to develop custom populations. Plant introductions (PI) are frequently used as source parents in mapping studies and for germplasm improvement, as was the case for mapping Phytophthora capsici resistance and developing resistant breeding lines (). Further, if a trait segregates closely with population structure, as was the case for seed size in C. pepo and maturity in C. moschata, this would indicate that populations should be formed by crossing between the groups identified to remove the confounding effects of population structure (). When higher density genotyping may be necessary or the PIs are not well characterized for a trait of interest, the data generated in this study can be used to prioritize accessions for re-sequencing and phenotyping. Our GenoCore analysis provides a subset of several hundred accessions that would likely be informative for re-sequencing efforts.

## CONFLICT OF INTEREST STATEMENT

Michael Mazourek is a co-founder of Row 7 Seeds, but neither receives compensation nor holds equity.

## AUTHOR CONTRIBUTIONS

COH wrote the first draft. MM, ZF, and RG provided project oversight. COH, JF, and KB conducted data analysis. KR, JL, and BJ assisted with data curation and germplasm selection. All authors contributed to the article.

## FUNDING

This work was supported by CucCAP, a USDA-NIFA-SCRI competitive grant 2015-51181-24285.

## ACKNOWLEDGMENTS

We thank Kyle LaPlant for plant phenotyping assistance, and Sue Hammer and Paige Reeves for assistance with DNA extraction.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study includeing variant and raw sequence data are available on the Cucurbit Genomics Dataase at cucurbitgenomics.org. The phenotypic data used are available for download from the USDA Germplasm Resources Information Network (GRIN; www.ars-grin.gov). Intermediate files and code used in the study are available on Github at www.github.com/ch728/Cucurbita-USDA.

## REFERENCES

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12. doi:10.1186/1471-2105-12-246

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

Chomicki, G., Schaefer, H., and Renner, S. S. (2020). Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. *New Phytologist* 226, 1240–1255. doi:10.1111/nph.16015

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10. doi:10.1093/gigascience/giab008

de los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: What is it? *PLOS Genetics* 11, e1005048. doi:10.1371/journal.pgen.1005048

Ding, W., Wang, Y., Qi, C., Luo, Y., Wang, C., Xu, W., et al. (2021). Fine mapping identified the gibberellin 2-oxidase gene CpDw leading to a dwarf phenotype in squash (cucurbita pepo l.). *Plant Science* 306, 110857. doi:10.1016/j.plantsci.2021.110857

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrBLUP. *The Plant Genome* 4, 250–255. doi:10.3835/plantgenome2011.08.0024

Ferriol, M. and Picó, B. (2008). Pumpkin and winter squash. In *Handbook of Plant Breeding* (Springer New York). 317–349. doi:10.1007/978-0-387-30443-4_10

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9, e90346. doi:10.1371/journal.pone.0090346

Gogarten, S. M., Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., et al. (2019). Genetic association testing using the GENESIS r/bioconductor package. *Bioinformatics* 35, 5346–5348. doi:10.1093/bioinformatics/btz567

Hernandez, C. O., Wyatt, L. E., and Mazourek, M. R. (2020). Genomic Prediction and Selection for Fruit Traits in Winter Squash. *G3 Genes|Genomes|Genetics* 10, 3601–3610. doi:10.1534/g3.120.401215

Jeong, S., Kim, J.-Y., Jeong, S.-C., Kang, S.-T., Moon, J.-K., and Kim, N. (2017). GenoCore: A simple and fast algorithm for core subset selection from large genotype datasets. *PLOS ONE* 12, e0181420. doi:10.1371/journal.pone.0181420

Kates, H. R., Soltis, P. S., and Soltis, D. E. (2017). Evolutionary and domestication history of cucurbita (pumpkin and squash) species inferred from 44 nuclear loci. *Molecular Phylogenetics and Evolution* 111, 98–109. doi:10.1016/j.ympev.2017.03.002

Kaźmińska, K., Hallmann, E., Rusaczonek, A., Korzeniewska, A., Sobczak, M., Filipczak, J., et al. (2018). Genetic mapping of ovary colour and quantitative trait loci for carotenoid content in the fruit of Cucurbita maxima Duchesne. *Molecular Breeding* 38, 114. doi:10.1007/s11032-018-0869-z

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Loy, J. B. (2004). Morpho-physiological aspects of productivity and quality in squash and pumpkins (cucurbita spp.). *Critical Reviews in Plant Sciences* 23, 337–363. doi:10.1080/07352680490490733

Lust, T. A. and Paris, H. S. (2016). Italian horticultural and culinary records of summer squash (cucurbita pepo, cucurbitaceae) and emergence of the zucchini in 19th-century milan. *Annals of Botany* 118, 53–69. doi:10.1093/aob/mcw080

McCouch, S., Navabi, Z. K., Abberton, M., Anglin, N. L., Barbieri, R. L., Baum, M., et al. (2020). Mobilizing crop biodiversity. *Molecular Plant* 13, 1341–1344. doi:10.1016/j.molp.2020.08.011

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., and Myles, S. (2015). LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. *G3 Genes|Genomes|Genetics* 5, 2383–2390. doi:10.1534/g3.115.021667

Montero-Pau, J., Blanca, J., Esteras, C., Martínez-Pérez, E. M., Gómez, P., Monforte, A. J., et al. (2017). An SNP-based saturated genetic map and QTL analysis of fruit-related traits in Zucchini using Genotyping-by-sequencing. *BMC Genomics* 18, 94. doi:10.1186/s12864-016-3439-y

Nee, M. (1990). The domestication ofcucurbita (cucurbitaceae). *Economic Botany* 44, 56–68. doi:10.1007/bf02860475

Paris, H. S. (2015). Germplasm enhancement of cucurbita pepo (pumpkin, squash, gourd: Cucurbitaceae): progress and challenges. *Euphytica* 208, 415–438. doi:10.1007/s10681-015-1605-y

Paris, H. S. and Brown, R. N. (2005). The genes of pumpkin and squash. *HortScience* 40, 1620–1630. doi:10.21273/hortsci.40.6.1620

Paris, H. S., Lebeda, A., Křistkova, E., Andres, T. C., and Nee, M. H. (2012). Parallel evolution under domestication and phenotypic differentiation of the cultivated subspecies of cucurbita pepo (cucurbitaceae). *Economic Botany* 66, 71–90. doi:10.1007/s12231-012-9186-3

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81, 559–575. doi:10.1086/519795

Savage, J. A., Haines, D. F., and Holbrook, M. N. (2015). The making of giant pumpkins: how selective breeding changed the phloem of cucurbita maxima from source to sink. *Plant, Cell & Environment* 38, 1543–1554. doi:10.1111/pce.12502

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., et al. (2017). Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes. *Molecular Plant* 10, 1293–1306. doi:10.1016/j.molp.2017.09.003

Wang, X., Ando, K., Wu, S., Reddy, U. K., Tamang, P., Bao, K., et al. (2021). Genetic characterization of melon accessions in the u.s. national plant germplasm system and construction of a melon core collection. *Molecular Horticulture* 1. doi:10.1186/s43897-021-00014-9

Wang, X., Bao, K., Reddy, U. K., Bai, Y., Hammar, S. A., Jiao, C., et al. (2018). The USDA cucumber (cucumis sativus l.) collection: genetic diversity, population structure, genome-wide association studies, and core collection development. *Horticulture Research* 5. doi:10.1038/s41438-018-0080-8

Wu, P.-Y., Tung, C.-W., Lee, C.-Y., and Liao, C.-T. (2019a). Genomic Prediction of Pumpkin Hybrid Performance. *The Plant Genome* 12, 180082. doi:10.3835/plantgenome2018.10.0082

Wu, S., Wang, X., Reddy, U., Sun, H., Bao, K., Gao, L., et al. (2019b). Genome of 'charleston gray', the principal american watermelon cultivar, and genetic characterization of 1,365 accessions in the u.s. national plant germplasm system watermelon collection. *Plant Biotechnology Journal* 17, 2246–2258. doi:10.1111/pbi.13136

Xanthopoulou, A., Montero Pau, J., Mellidou, I., Kissoudis, C., Blanca, J., Picó, B., et al. (2019). Whole-genome resequencing of Cucurbita pepo morphotypes to discover genomic variants associated with morphology and horticulturally valuable traits. *Horticulture Research* 6, 94. doi:10.1038/s41438-019-0176-9

Xiang, C., Duan, Y., Li, H., Ma, W., Huang, S., Sui, X., et al. (2018). A high-density EST-SSR-based genetic map and QTL analysis of dwarf trait in cucurbita pepo l. *International Journal of Molecular Sciences* 19, 3140. doi:10.3390/ijms19103140

Yu, J., Wu, S., Sun, H., Wang, X., Tang, X., Guo, S., et al. (2022). CuGenDBv2: an updated database for cucurbit genomics. *Nucleic Acids Research* doi:10.1093/nar/gkac921

Zhang, G., Ren, Y., Sun, H., Guo, S., Zhang, F., Zhang, J., et al. (2015). A high-density genetic map for anchoring genome sequences and identifying QTLs associated with dwarf vine in pumpkin (Cucurbita maxima Duch.). *BMC Genomics* 16, 1101. doi:10.1186/s12864-015-2312-8

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi:10.1093/bioinformatics/bts606

Zhong, Y.-J., Zhou, Y.-Y., Li, J.-X., Yu, T., Wu, T.-Q., Luo, J.-N., et al. (2017). A high-density linkage map and QTL mapping of fruit-related traits in pumpkin (Cucurbita moschata Duch.). *Scientific Reports* 7, 12785. doi:10.1038/s41598-017-13216-3

## FIGURE CAPTIONS

**Figure 1.** Cross-validation error plots used to pick the optimum K value for admixture analysis. The K value that balances minimizing cross-validation error and parsimony, and thus chosen for the final analysis, is labeled with a red point.

**Figure 2.** Geographical distribution of the USDA Cucurbita ssp. collection. The size of the pie chart is scaled according to the number of accessions and sector areas correspond to the proportion of the three species.

**Figure 3.** Spatial distribution of filtered markers for **panel a.** *C. pepo*, **panel b.** *C. moschata*, and **panel c.** *C. maxima.*

**Figure 4.** Population structure results aligned vertically by species. **Panel a.** Admixture plots: each stacked barplot represents an accession colored by proportion of inferred ancestral population. Groups based on hierarchical clustering are delimited by vertical bars and labeled with numbers along the bottom. **Panel b.** Plots of the first two principal components (PC) of accessions colored by region, variation explained by PCs is labeled on each axis.

**Figure 5.** Ancestry coefficients projected on cultivars from each species. Results are shown grouped by market/varietal class.

**Figure 6.** Curves showing R-squared value as a measure of LD on the y-axis and distance between markers in megabases on the x-axis.

**Figure 7.** Histograms of continuous and ordinal traits for **panel a.** *C. pepo*, **panel b.** *C. moschata*, and **panel c.** *C. maxima.*

**Figure 8.** GWAS results: **panel a.** *C. pepo* plant_type (bush or vine historical data, Bu gene), plant_type2( bush or vine contemporary data, Bu), flesh_max (flesh maximum thickness), gn_fruit (green fruit color), or_flesh (orange flesh color); **panel b.** *C. moschata* fruit_len (fruit length); **panel c.** *C. maxima* rib (degree of fruit ribbing); gn_fruit (green fruit color)

**Figure 9.** Quantile-quantile plots matching GWAS results. **Panel a.** *C. pepo*, **panel b.** *C moschata*, and **panel c**. *C. maxima*

## TABLES

| | *C. pepo* | | *C. moschata* | | *C. maxima* | |
|---|---|---|---|---|---|---|
| Chromosome | Raw | Filtered | Raw | Filtered | Raw | Filtered |
| 0 | 12498 | 2550 | 2708 | 546 | 1501 | 132 |
| 1 | 7497 | 2831 | 3890 | 1468 | 4185 | 121 |
| 2 | 5153 | 2049 | 3661 | 1538 | 2101 | 55 |
| 3 | 4875 | 1943 | 3472 | 1499 | 2201 | 51 |
| 4 | 4598 | 1982 | 6880 | 2553 | 5703 | 106 |
| 5 | 4045 | 1628 | 2716 | 887 | 3115 | 46 |
| 6 | 3871 | 1384 | 3262 | 1159 | 3035 | 92 |
| 7 | 3129 | 1222 | 2668 | 969 | 2705 | 62 |
| 8 | 3875 | 1583 | 2348 | 810 | 2391 | 61 |
| 9 | 3766 | 1390 | 3106 | 995 | 2750 | 84 |
| 10 | 3585 | 1488 | 3550 | 1327 | 2297 | 52 |
| 11 | 3227 | 1216 | 4336 | 1830 | 3713 | 131 |
| 12 | 3089 | 1163 | 3711 | 1330 | 2026 | 47 |
| 13 | 3434 | 1350 | 3106 | 1280 | 2131 | 82 |
| 14 | 3543 | 1291 | 4753 | 1929 | 4317 | 100 |
| 15 | 2640 | 960 | 3564 | 1321 | 2662 | 58 |
| 16 | 3088 | 1060 | 2933 | 1107 | 2058 | 100 |
| 17 | 2994 | 1175 | 2885 | 1096 | 2195 | 86 |
| 18 | 3053 | 1258 | 3341 | 1316 | 1826 | 46 |
| 19 | 3381 | 1340 | 2638 | 990 | 1793 | 46 |
| 20 | 3096 | 1155 | 2497 | 903 | 1893 | 41 |
| Total | 88437 | 32018 | 72025 | 26853 | 56598 | 1599 |

**Table 1.** Distribution and number of raw and filtered SNPs per chromosome for each species

**Figure 10.** Results from running GenoCore in each of the panels. **Panel a** shows the PCA plots for each panel with accessions selected by GenoCore represented as black points.**Panel b** shows the proportion of total accessions needed to obtain a certain coverage of diversity

| | *C. pepo* | *C. moschata* | *C. maxima* |
|---|---|---|---|
| 1 | Mixed Group; Many from Spain, Turkey, and Syria | Mostly from Mexico | Mixed; Primarily from South America and Asia |
| 2 | Wild subsp. *ovifera* var. *texana* and var. *ozarkana*; North American | Mostly Mexico and Guatemala | Mixed; Primarily from Asia and Europe |
| 3 | Majority from Turkey | Mostly from Mexico | Mostly from North Macedonia |
| 4 | Majority from North Macedonia | Mostly from Africa | Mostly from Argentina |
| 5 | Majority from Egypt | Mostly from India | Mostly Turkey, Iran, Afghanistan |
| 6 | Majority from Mexico | Mixed origin Europe and Americas;Many similar to cheese or neck type | Mostly from Africa |
| 7 | Majority from Syria | | |
| 8 | Majority from Pakistan and Afghanistan | | |
| 9 | Majority from Spain | | |
| 10 | Wild subsp. *fraterna*; Central American | | |

**Table 2.** Commonalities among accessions in each group, most groupings are dictated by geography

| **Species** | **Trait** | **Type** | **Description** | $h^2_G$ |
|---|---|---|---|---|
| *C. pepo* | | | | |
| | seed_wt | Quantitative | Weight of 100 seeds in grams | |
| | plant_type | Binary | Historical plant architecture data coded as vining or bush | NA |
| | plant_type2 | Binary | Contemporary plant architecture data coded as vining or bush | NA |
| | max_vig | Ordinal | Maximum plant vigor on 1-5 scale | |
| | min_vig | Ordinal | Minimum plant vigor on 1-5 scale | |
| | max_width | Quantitative | Maximum fruit width in centimeters | |
| | width_min | Quantitative | Minimum fruit width in centimeters | |

| | | | | |
|---|---|---|---|---|
| | len_max | Quantitative | Maximum fruit length in centimeters | |
| | len_min | Quantitative | Minimum fruit length in centimeters | |
| | flesh_max | Ordinal | Maximum fruit thickness in centimeters | |
| | sb_nymph | Quantitative | Number of squash bug nymphs on plant | |
| | sb_adult | Quantitative | Number of adult squash bugs on plant | |
| | cuc_inj | Ordinal | Severity of beetle damage on a 0-4 scale | |
| | or_flesh | Binary | Flesh color coded as orange or not orange | NA |
| | yl_flesh | Binary | Flesh color coded as yellow or not yellow | NA |
| | yl_fruit | Binary | Color of fruit coded as yellow or not yellow | NA |
| | tan_fruit | Binary | Color of fruit coded as tan or not tan | NA |
| | gn_fruit | Binary | Color of fruit coded as green or not green | NA |
| | globe_fruit | Binary | Fruit shape as globe or not globe | NA |
| | oblong_fruit | Binary | Fruit shape as oblong or not oblong | NA |
| | smooth_fruit | Binary | Furit texture as smooth or not smooth | NA |
| | rib_fruit | Binary | Degree of ribbing | NA |
| | spec_fruit | Binary | Fruit patterning as speckled or not speckled | NA |
| | mot_fruit | Binary | Fuit patterning as mottled or not mottled | NA |
| | solid_fruit | Binary | Fruit patterning as solid color or patterned | NA |
| *C. moschata* | | | | |
| | fruit_len | Quantitative | Fruit length in centimeters | |
| | fruit_diam | Quantitative | Fruit diameter in centimeters | |
| | maturity | Binary | Fruit maturity as early or late | NA |
| | or_fruit | Binary | Fruit color coded as orange or not orange | NA |

| | | | | |
|---|---|---|---|---|
| | smooth_fruit | Binary | Fruit surface texture encoded as smooth or not smooth | NA |
| *C. maxima* | | | | |
| | len | Quantitative | Fruit length in centimeters | |
| | set | Ordinal | Fruit set from poor to excellent (1-9) | |
| | diam | Quantitative | Fruit diameter in centimeters | |
| | watermelon_mosaic | Ordinal | Susceptibility to WMV from slight to severe (0-9) | |
| | cuc_mosaic | Ordinal | Cucumber mosaic susceptibility from slight to severe (0-9) | |
| | maturity | Quantitative | Number of days from field transplanting to date of first pollination | |
| | unif | Ordinal | Fruit uniformity from poor to excellent (1-9) | |
| | pm | Ordinal | Susceptibility to PM from slight to severe (0-9) | |
| | plant_habit | Binary | Plant type as vining or not vining | NA |
| | vig | Ordinal | Plant vigor from poor to excellent (1-9) | NA |
| | or_flesh | Binary | Flesh color as orange or not orange | NA |
| | rib | Ordinal | Fruit ribbing from slight to pronounced (1-9) | |
| | fruit_spot | Ordinal | Fruit spotting from slight to pronounced (1-9) | |
| | gray_fruit | Binary | Fruit color encoded as gray or not gray | NA |
| | or_fruit | Binary | Fruit color encoded as orange or not orange | NA |
| | gn_fruit | Binary | Fruit color encoded as green or not green | NA |

Table 3 Commonalities among accessions in each group, most groupings are dictated by geography