Characterization of the USDA Cucurbita pepo, Cucurbita moschata, and Cucurbita maxima Collections

This manuscript (permalink) was automatically generated from ch728/cucurbit-usda@c6f3e0b on August 30, 2021.

Authors

- **Christopher Owen Hernandez**
  0000-0002-1668-7121
  Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY
- **Jack Fabrizio**

  Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY
- **Joanne Labate**

  USDA, Geneva, NY
- **Zhangjun Fei**
  0000-0001-9684-1450
  Boyce Thompson Institute for Plant Research, Ithaca, NY
- **Michael Mazourek**
  0000-0002-2285-7692
  Department of Plant Breeding and Genetics, Cornell University, Ithaca,NY

Abstract

The *Cucurbita* genus is home to a number of economically and culturally important species. We present the analysis of genotyping-by-sequencing data generated from sequencing the USDA germplasm collections of *Cucurbita pepo*, *Cucurbita moschata*, and *Cucurbita maxima*. These collections include a mixture of wild, landrace, and cultivated specimens from all over the world. Roughly 4,000 - 40,000 quality 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. GWAS was conducted for each data set using both historical and contemporary data, and signals were detected for several traits, including the bush gene (*Bu*) in *C. pepo*. These data represent the largest collection of sequence *Cucurbita* and can be used to direct the maintenance of genetic diversity, develop 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*.

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 sativa*), melon (*Cucumis melo*), watermelon (*Citrullus lanatus*), and squash (*Cucurbita ssp.*) [1]. Like other cucurbits, squash exhibit diversity in growth habit, fruit morphology, metabolite content, disease resistance, and have a nuanced domestication story [2,3]. The genomes of *Cucurbita ssp.* are small (roughly 500 Mb), but result from complex interactions between ancient genomes brought together through an allopolyploidization event [4]. 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 broadlycultivated: *Cucurbita maxima*, *Cucurbita moschata*, and *Cucurbita pepo* [1]. Few genomic resources have been available for working with these species; although, draft genomes and annotations, along with web-based tools and other genomics data are emerging [5]. 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 [6,7,8,9,10,11]; however, there has yet to be a comprehensive survey of the genetic diversity in large diverse *Cucurbita* germplasm panels, such as those maintained by the USDA within the Germplasm Resources Information Network (GRIN) 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 [12]. 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 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 [13,14]. 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 [15]. 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 [16]. Many squash are known as pumpkins; the pumpkin designation is aculture 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 [1]. Cultivars deemed as pumpkins can be found in all widely cultivated squash species. Unlike the previous groupings, morophotypes/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 [17].

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* [10]. 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* [17]. Europe played a crucial role as a secondary center of diversification for *pepo*, but not *ovifera* [18]. 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 [19].

The origin of *C. moschata* is more uncertain than *C. pepo*; it is unclear whether *C. moschata* has a South or North American origin [3]. 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 [4]. *C. moschata* plays an important role in squash breeding as it cross-fertile to various degrees with *C. pepo* and *C. maxima*, and can thus be used as a bridge to move genes across species [4]. 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 [1].

C. maxima contains many popular winter squash including Buttercup/Kobocha types, Kuri, Hubbard, and Banana squash [1]. 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 [20]. 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 [17]. *C. maxima* is believed to have a South American origin, and was likely domesticated near Peru, with a secondary center of domestication in Japan/China [nee_domestication_1990; [4]].

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 data from each of these collections, population genomics analysis, results from genome-wide association using historical and contemporary phenotypes, and develop a core panel for re-sequencing.

Material and Methods

Plant Materials and Genotyping

All available germplasm were requested from USDA cooperators for *C. maxima* (534), *C. moschata* (314), and *C. pepo* (829) 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.

Variant Calling and Filtering

SNP calling was conducted using the TASSEL-GBS V5 pipeline [21]. Tags produced by this pipeline were aligned using the default settings of the BWA aligner [22]. Raw variants were filtered using VCFtools [23]. Before filtering SNPs, samples with a total read depth of ≥ 2 standard deviations below the mean of all samples were removed before further analysis. Settings for filtering SNPs were as follows, minor allele frequency (MAF) ≥ 0.01, missingness ≤ 0.5, and biallelic. Three outlier genotypes were found in an initial PCA analysis 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.

Population Genomics Analysis

ADMIXTURE [24], 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; thus, marker sets were further filtered to obtain SNPs in approximate linkage equilibrium using the "–indep-pairwise" option in PLINK [25] with r^2 set to 0.1, a window size of 50 SNPs, and a 10 SNP step size . All samples labeled as cultivars were removed from the data prior to running ADMIXTURE. Cross-validation was used to determine the best k value for each species. Briefly, ADMIXTURE was run with different k 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 (Supplemental Figures Figure ??. 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. he original filtered marker data, not the LD-pruned data used for ADMIXTURE, were converted to a dosage matrix using VCFtool's "–012" argument. A kinship matrix $\mathbf{K}$ was created using the dosage matrix as input to the "A.mat()" function in Sommer [26]. PCA was conducted using the R function "princomp()" with $\mathbf{K}$ supplied as the covariance matrix.

Phylogenetic analysis was conducted in a subset of the *C. pepo* panel with clearly labeled subspecies information or where enough information to unambiguously assign the accession to a subspecies was present. The SNPhylo [27] pipeline was used to infer an unrooted tree using the maximum likelihood method. Default settings were used, except the minimum coverage parameter was decreased to 3 instead of 5 to account for the lower average coverage of GBS data.

Analysis of Phenotypic Data

Historical data were obtained from the USDA Germplasm Resources Information Network (GRIN; http://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. 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.

Genomic heritability [28] (h_g^2) was calculated for all phenotypes. The parameter h_g^2 was calculated for continuous traits using the formula $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$, where σ_g^2 and σ_e^2 are genetic and error variances estimated from a whole-genome regression of phenotype on marker data using ASReml-R . Multi-class categorical traits were converted to one or several different binary traits depending on the number of entries in each category. For binary traits, a Logit model was fit for the binary response and the heritability was estimated as $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2+\frac{\pi^2}{3}}$ [29]. In addition to heritability, the amount of phenotypic variance explained by population structure (R^2_{pop}) was calculated from a multiple linear regression of phenotype on sturcture inferred by ADMIXTURE. The R function lm was used to regress continuous phenotypes on the $\mathbf{Q}$ matrix obtained from ADMIXTURE. The R glm function was used with "family=binomial" to regress binary traits on population structure. As there is no R^2 defined for logistic models, McFadden's psuedo R^2 was used to assess the correlation between binary traits and population structure [30].

GWAS

Data were imputed prior to association analysis. LinkImpute [31], as implemented by the TASSEL [32] "LDKNNiImputatioHetV2Plugin" plugin was used for imputation with default settings. Any data still missing after this process were mean imputed. The GENESIS [**doi?** 10.1093/bioinformatics/btz567] R package, which can model both binary and continuous traits, was used for association. 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 ($\mathbf{K}$) matrix ($u \sim N(0, \sigma_u^2\mathbf{K})$). Binary traits were modeled using the logistic regression feature in GENESIS.

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

Creation of a Core Collection

Subsets representative of each panel's genetic diversity were identified through running GenoCore [33] on each of the filtered SNP sets. A subset of the *C. pepo* panel and key genotypes from the other two species were combined to form a core collection for the cucurbit community. Key genotypes were chosen to represent important market classes and for variation based on variation in traits. These genotypes will be further purified through two additional rounds of selfing and then resequenced using skim-sequecing to produce whole-genome data.

Results

Genotyping

Each *Cucurbita ssp.* collection was genotyped using the Cornell Genotype by Sequencing (GBS) protocol. This resulted in 534 accessions for *C. maxima*, 314 for *C. moschata*, and 829 for *C. pepo*. Figure 1 shows the regional 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 (108,279) followed by *C. moschata* (85,345) and *C. maxima* (56,598). After filtering, *C. pepo* and *C. moschata* had a similar number of SNPs, around 40,000, whereas *C. maxima* had an order of magnitude fewer filtered SNPs (4787). This discrepancy may be an artifact of using Pst1, a rarer base-cutter previously optimized for use in *C. maxima* [34], rather than ApeK1 which was used for *C.pepo*

and *C. moschata*. The number and distribution of SNPs across each chromosomes is shown in Table 1.

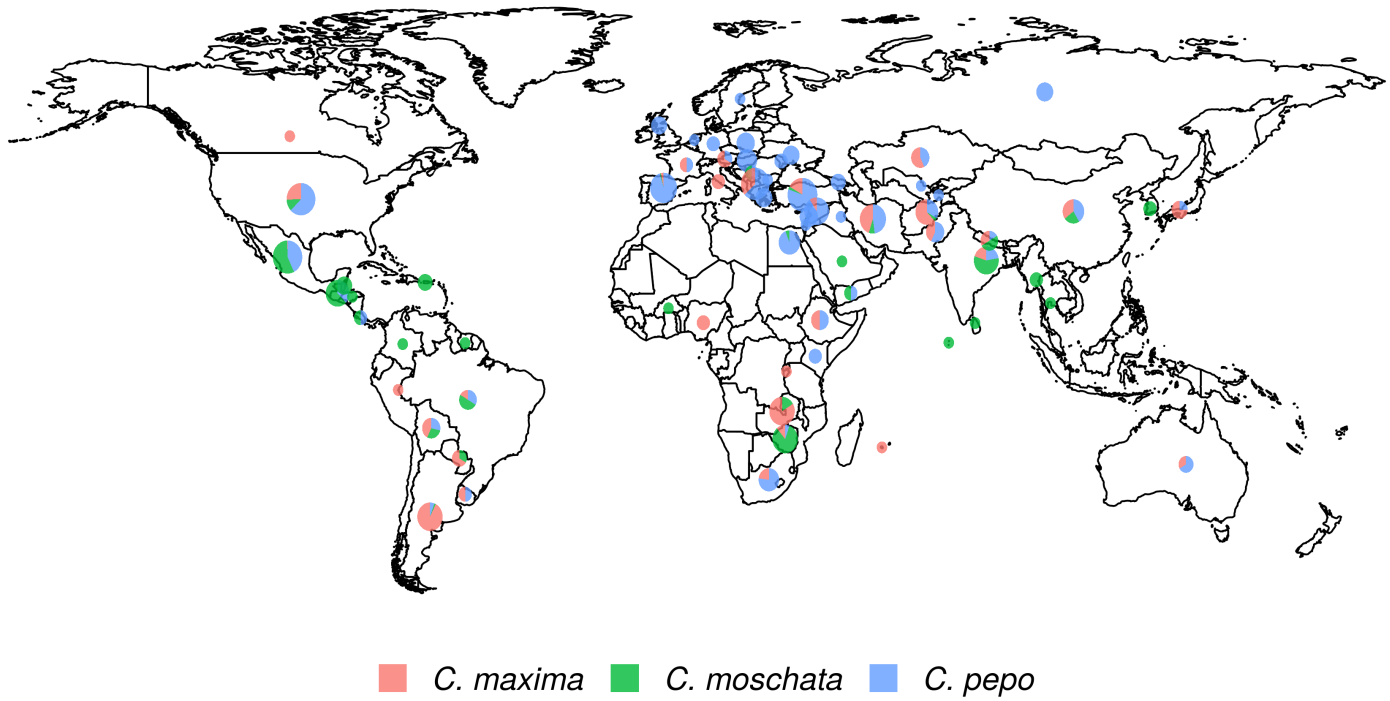


Figure 1: Geographical distribution of the USDA Cucurbita ssp. collection. The size of the pie chart is scaled according to the number of accessions and sector areas correspond to the proportion of the three species.

Table 1: Distribution and number of raw and filetered SNPs per chromosome for each species

Chrom.	***C. pepo***		***C. moschata***		***C. maxima***	
	Raw	Filtered	Raw	Filtered	Raw	Filtered
0	16901	5656	3748	1236	1501	419
1	9245	4155	4575	2627	4185	300
2	6160	2921	4092	2535	2101	169
3	5908	2668	3815	2393	2201	157
4	5540	2652	7868	4458	5703	382
5	4813	2254	3226	1804	3115	154
6	4555	2100	3663	2182	3035	345
7	3677	1761	3300	1784	2705	148
8	4551	2189	2692	1577	2391	191
9	4521	1995	3427	1902	2750	229
10	4366	2052	4219	2225	2297	120
11	3839	1727	5212	2962	3713	309
12	3777	1614	5329	2286	2026	162
13	4002	1879	3888	2013	2131	257
14	4275	1973	5568	3198	4317	297
15	3086	1427	3911	2358	2662	172
16	4274	1589	3407	1987	2058	302

Chrom.	*C. pepo*		*C. moschata*		*C. maxima*	
17	3519	1657	3557	1888	2195	251
18	3568	1723	3775	2105	1826	133
19	4015	1860	3278	1716	1793	169
20	3687	1692	3795	1623	1893	133
Total	108279	47544	85345	46859	56598	4799

Population Structure and Genetic Diversity

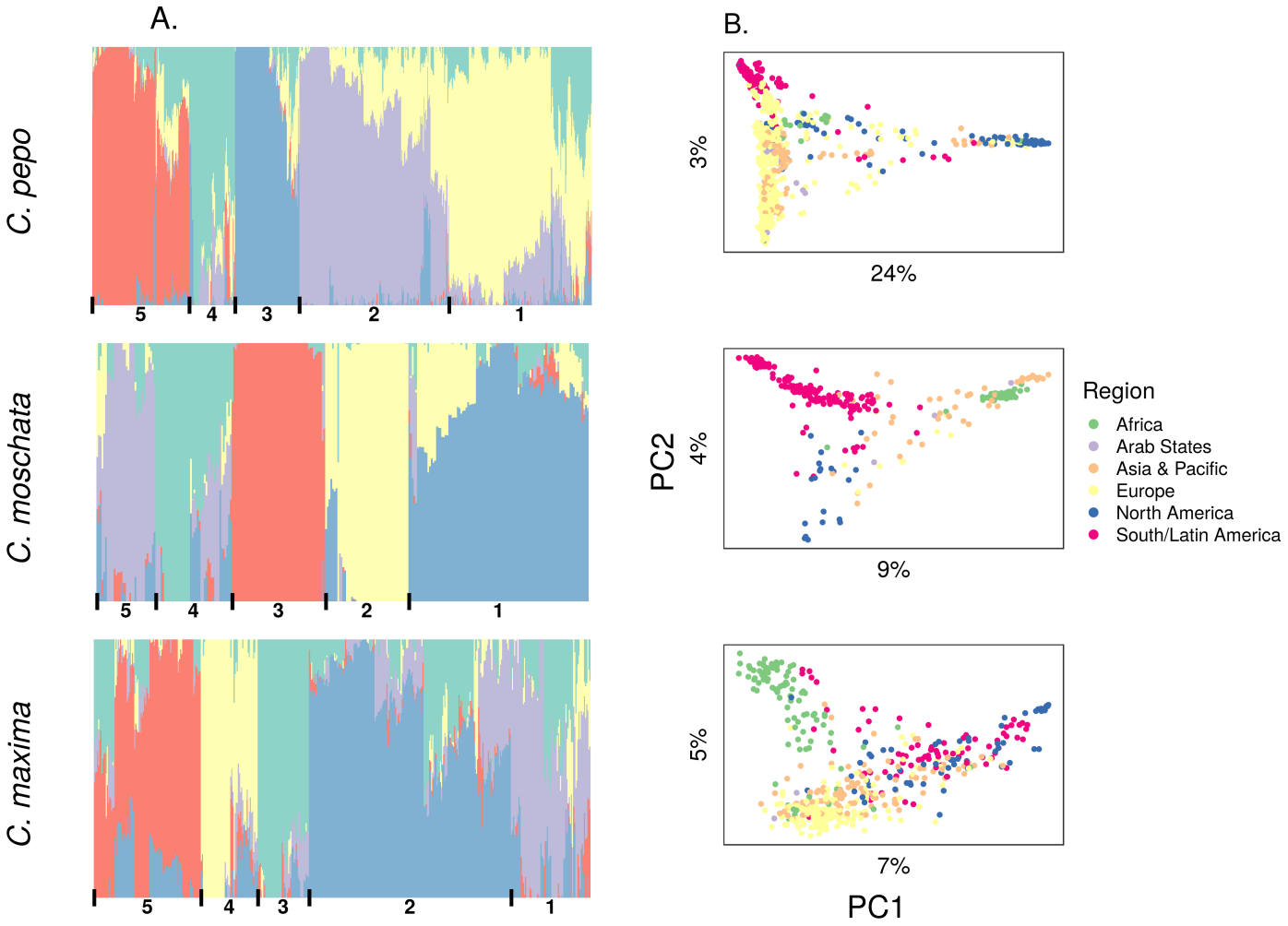


Figure 2: Population structure results aligned vertically by species. (A) Admixture plots: each stacked barplot represents an accession colored by proportion of inferred ancestral population. Groups based on hierarchical clustering are delimited by vertical bars and labeled with numbers along the bottom. (B) Plots of the first two principle components (PC) of accessions colored by region, variation explained by PCs is labeled on each axis.

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

Group	Species		
	C. pepo	*C. moschata*	*C. maxima*
1	Europe/Asia, mostly for Turkey	South American/Latin American	Mixed origin; kobocha/turban types
2	Europe, mostly from Macedonia	South American/Latin American	European, mostly from Macedonia

Group	Species		
3	North America, wild and landrace *ovifera*	African	Asia
4	Mixed origin	India	South American
5	South America, mostly from Mexico	Mixed origin; elongated fruit type	African

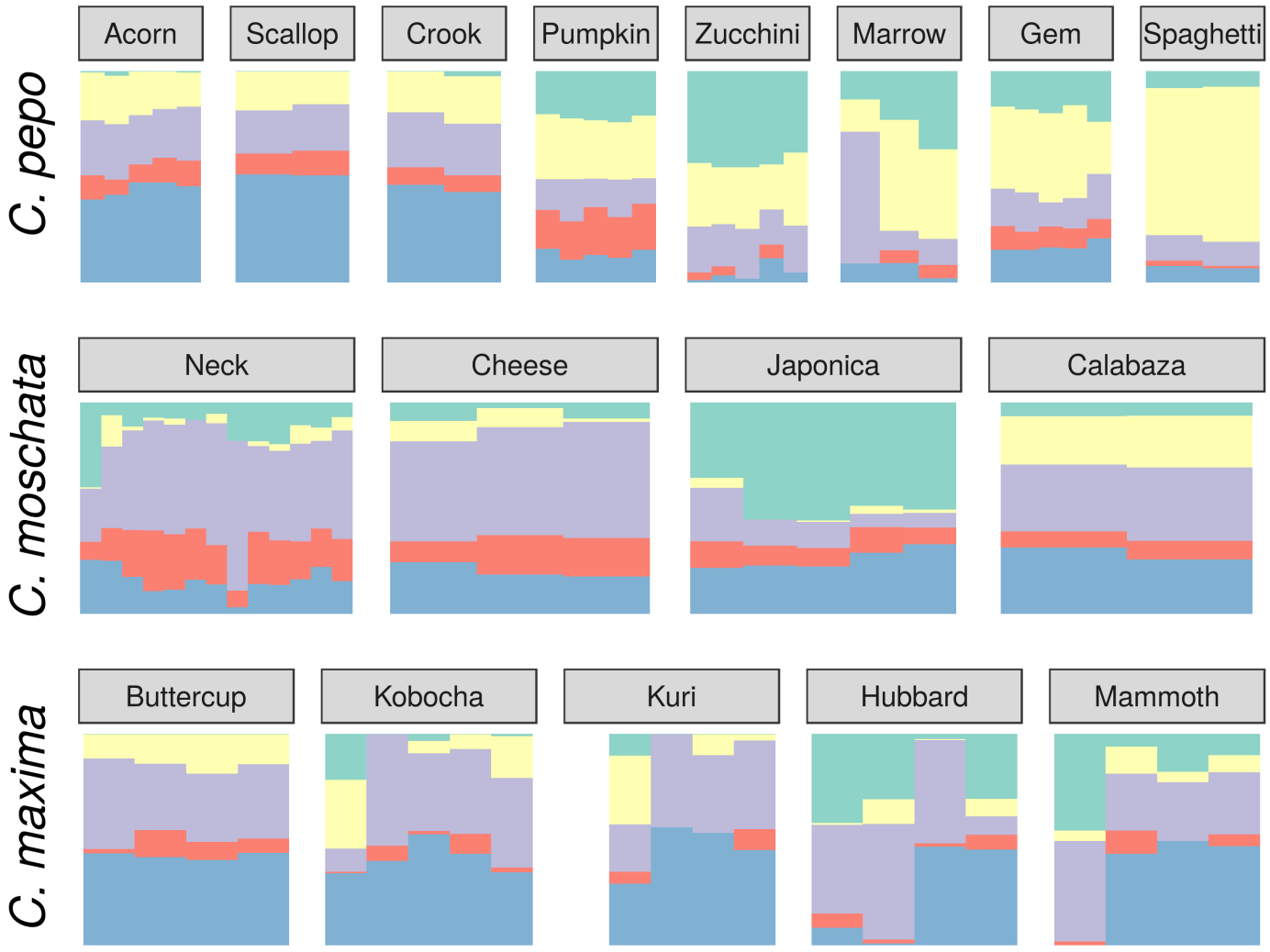


Figure 3: Ancestry coefficients projected on cultivars from each species. Results are shown grouped by market/varietal class.

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 2 panel A. These data support five ancestral groups (K=5) in each of the species. 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) derived from principal components analysis (PCA) of the marker data are shown in Figure 2 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*.

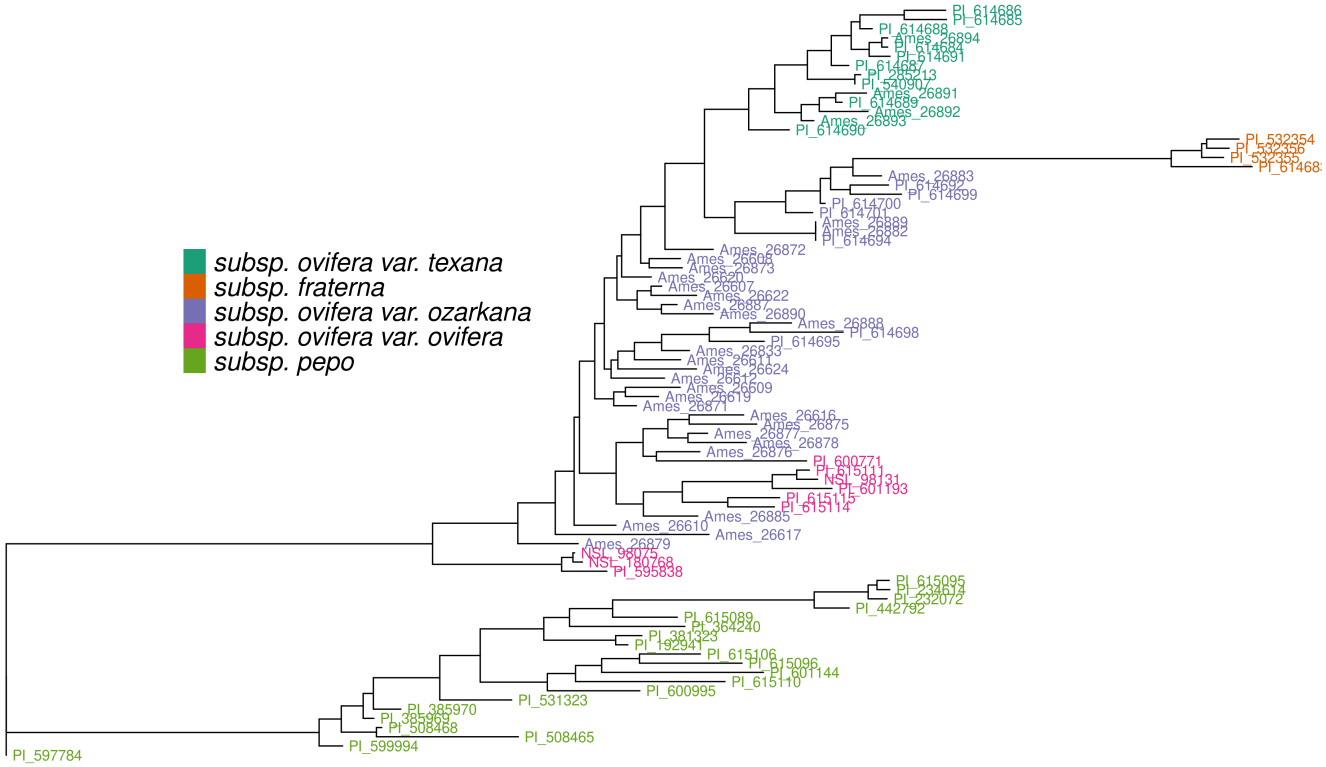


Figure 4: Unrooted maximum likelihood tree of *C. pepo* subspecies inferred using wild and cultivated germplasm in the *C. pepo* collection.

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 (jacko' lantern), Zuchinni, Marrow, Gem, and Spaghetti; Neck, Cheese, Japonica, and Calabaza in *C. moschata*; and Buttercup, Kobocha, Kuri, Hubbard, and Mammoth (show squash) in *C. maxima*. These groupings are shown in Figure 3. 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 50% 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 and Cheese type market classes showed very similar ancestry patterns, whereas the Japonica and Calabaza types were more distinct. Relative to the *C. pepo* and *C. moschata*, the *C. maxima* cultivars were less distinct from one another.

Unlike *C. moschata* and *C. maxima*, several different subspecies were present in the *C. pepo* collection, including some wild specimens. A group of 82 accessions with 11,065 high-quality SNPs was used for constructing an unrooted phylogenetic tree. The tree is shown in Figure 4 and recapitulates the relationships among *pepo* subspecies shown in previous work [17].

Analysis of Phenotypic Data

All available historical data from GRIN were compiled. Only traits with $\geq$ 100 entries were considered for further analysis. Filtering resulted in 21 traits for *C. pepo*, 5 for *C. moscahta* 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 $\sim$ 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*.

Around half of the traits were quantitative/ordinal and half were categorical and coded as binary traits, see Table 3. The majority of traits measured on a quantitative scale were normally distributed. Marker-based narrow-sense heritability (h^2_G) was calculated for each trait. Values for h^2_G ranged from 0.12 to close to 1. Most traits had moderate to high heritabilities ($\geq$ 0.4). Regression of trait data on the $\mathbf{Q}$ matrix obtained from structure analysis was used to determine the amount of phenotypic variation explained by population structure. In *C. pepo*, traits related to fruit morphology tended to have high correlations with population structure (R^2_{pop}). Seed weight had the highest correlation with an R^2_{pop} of 0.6. In *C. moschata*, maturity showed the highest correlation with population structure (R^2_{pop} of 0.52). None of the 16 traits in *C. maxima* had a high correlation with population structure. The only exception was plant growth habit.Traits related to pest resistance were measure in *C. maxima* and *C. pepo* and had among the lowest correlations with population structure.

Table 3: Summary of trait data used for association analysis. A brief description is given for each trait followed by the number of entries with records, the genomic heritability (h^2_G), and the R-square value from multiple linear regression of population structure on phenotype (R^2_{pop}). Trait names followed by a number (e.g. Fruit Color1) are traits derived from alternative coding of multi-class categorical traits.

	Trait	Description	Pop Size	h^2_G	R^2_{pop}
C. pepo					
	Max Fruit Thickness	Maximum fruit thickness in centimeters	421	0.72	0.27
	Min Fruit Thickness	Minimum fruit thickness in centimeters	174	0.58	0.14
	Min Fruit Length	Minimum fruit length in centimeters	413	0.82	0.37
	Max Fruit Length	Maximum fruit length in centimeters	315	0.91	0.33
	Max Fruit Width	Minimum fruit width in centimeters	303	1.00	0.49
	Min Fruit Width	Maximum fruit width in centimeters	413	1.00	0.49
	Fruit Texture	Fruit texture coded as smooth or not smooth	130	0.55	0.23
	Fruit Skin Pattern	Skin patterning coded as solid color or patterned	248	0.58	0.27
	Fruit Shape1	Fruit shape coded as oblong or not oblong	331	0.69	0.60

	Trait	Description	Pop Size		
	Fruit Shape2	Fruit shape coded as globe or not globe	331	0.67	0.48
	Flesh Color	Flesh color coded as either yellow or orange	377	0.53	0.19
	Fruit Color1	Color of fruit coded as yellow or not yellow	181	0.55	0.19
	Fruit Color2	Color of fruit coded as green or not green	181	0.68	0.55
	Cucumber Beetle Damage	Severity of beetle damage on a 0-4 scale	248	0.32	0.08
	Adult Squash Bug	Number of adult squash bugs on plant	237	0.88	0.07
	Nymph Squash Bug	Number of squash bug nymphs on plant	166	0.46	0.02
	Plant Type1	Historical plant architecture data coded as vining or bush	404	0.64	0.37
	Plant Type2	Contemporary plant architecture data coded as vining or bush	293	1.00	0.36
	Plant Vigor1	Minimum plant vigor on 1-5 scale	414	0.54	0.14
	Plant Vigor2	Maximum plant vigor on 1-5 scale	414	0.54	0.14
	100 Seed Wt.	Weight of 100 seeds in grams	822	0.90	0.60
C. moscahta					
	Fruit Color	Fruit color coded as orange or not orange	140	0.43	0.13
	Fruit Surface Texture	Fruit surface texture encoded as smooth or not smooth	127	0.18	0.07
	Fruit Diameter	Fruit diameter in centimeters	122	0.62	0.18
	Fruit Length	Fruit length in centimeters	121	1.00	0.18

	Trait	Description	Pop Size		
	Maturity	Fruit maturity on scale of early to late (1-8)	108	1.00	0.52
C. maxima					
	Fruit Color1	Fruit color encoded as gray or not gray	183	0.53	0.17
	Fruit Color2	Fruit color encoded as orange or not orange	183	0.57	0.08
	Fruit Color3	Fruit color encoded as green or not green	183	0.46	0.15
	Flesh Color	Flesh color on a scale of yellow to dark orange (1-5)	231	0.44	0.09
	Flesh Depth	Flesh thickness in centimeters	251	0.29	0.01
	Fruit Diameter	Fruit diameter in centimeters	248	0.37	0.29
	Fruit Length	Fruit length in centimeters	248	0.49	0.27
	Fruit Spot	Fruit spotting from slight to pronounced (1-9)	193	0.40	0.01
	Fruit Ribbing	Fruit ribbing from slight to pronounced (1-9)	243	0.64	0.14
	Powdery Mildew Susceptibility	Susceptibility to PM from slight to severe (0-9)	211	0.33	0.06
	WMV Susceptibility	Susceptibility to WMV from slight to severe (0-9)	212	0.19	0.05
	Fruit Set	Fruit set from poor to excellent (1-9)	251	0.36	0.15
	Uniformity	Fruit uniformity from poor to excellent (1-9)	244	0.35	0.07
	Vigor	Plant vigor from poor to excellent (1-9)	251	0.12	0.00
	Plant Type	Plant type as vining or not vining	251	0.74	1.19

	Trait	Description	Pop Size		
	Days to Pollen	Number of days from field transplanting to date of first pollination	236	0.52	0.15

Genome-wide Association

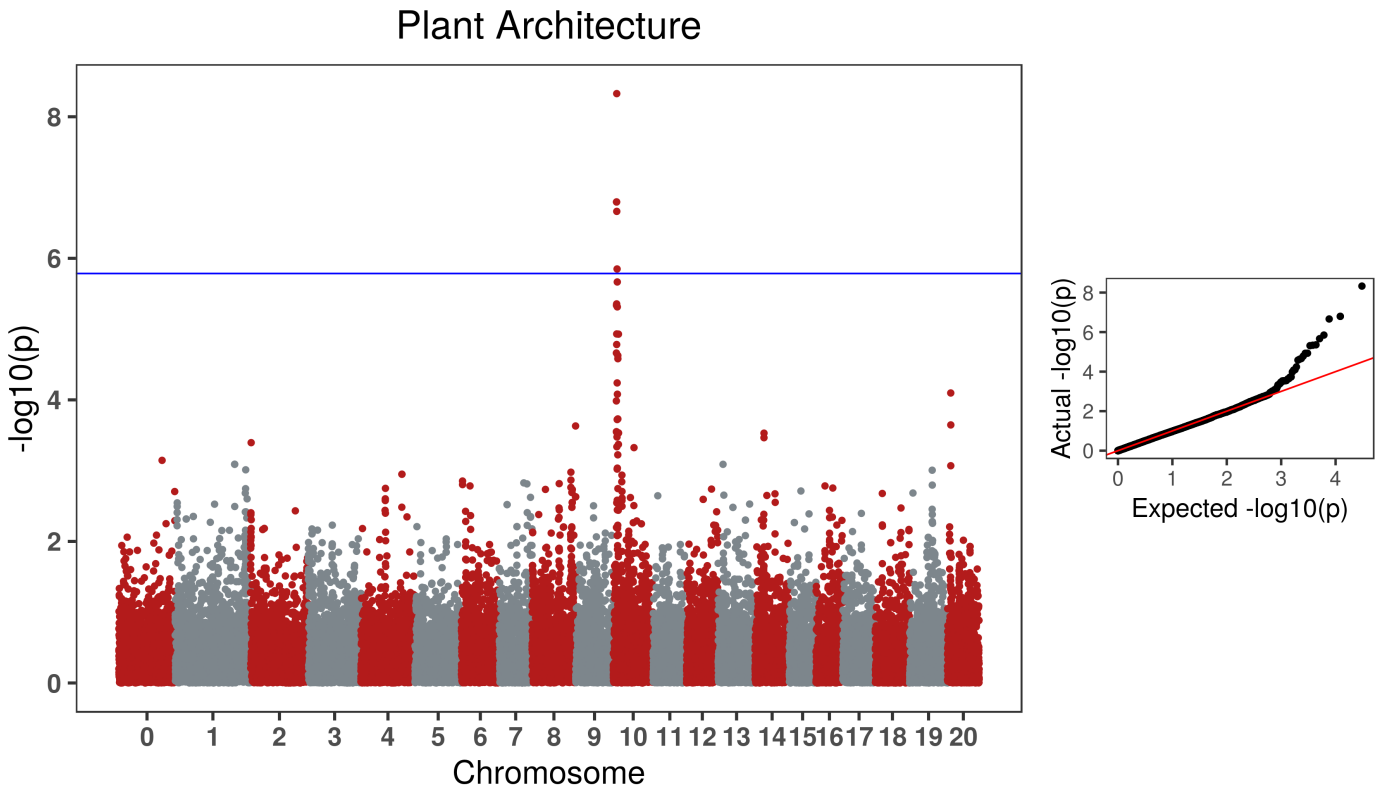


Figure 5: GWAS result for the Bush gene (*Bu*) in *C. pepo*

Genome-wide association was conducted for all traits using standard mixed-model analysis. No significant signals were detected in *C. moschata*. A weak signal was detected in *C. maxima* for fruit set on chromosome 12 and fruit ribbing on chromosome 17. Three phenotypes were significantly associated with SNPs in *C. pepo*: bush/vine plant architecture on chromosome 10, fruit flesh color on chromosome 5, and fruit width on chromosome 3. The bush/vine phenotype exhibited the strongest signal, and the Manhatten plot and p-value quantile-quantile plot is shown in Figure 5.

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

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 10 to 20% of the accessions were required to capture the genetic diversity in the panels (See Supplemental Figures). This amounted to 245 accessions in *C. pepo*, 154 in *C. moschata*, and in 248 *C.maxima*. The core subset identified in *C. pepo* was augmented with accessions that represented key market classes or that had traits of interest to breeding programs. Additionally, key accessions were selected from *C. maxima*, *C. moschata* and some wild species. Together these genotypes were purified through two additional rounds of selfing and seed will serve as the basis for a *Cucurbita ssp.* core to be used by breeding programs and researchers for further studies.

Discussion

Cucurbita pepo, *Cucurbita moschata*, and *Cucurbita 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 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 *Cucurbita 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.
Phylogenetic anlaysis of *Cucurbita pepo* using the whole-genome markers also supported the known relationships between the various subspecies in *pepo*. Together with the mapping of a putative bush gene (*Bu*) that appears to be syntenic with the bush gene mapped in *C. maxima*, we have demonstrated that these data constitute a new, high quality genetic resource for the Cucurbit community. These markers and our analysis of available germplasm have a number of uses for breeding and future experments aimed at biological insight.

our data provides many genome-wise markers which could be used to develop marker panels for use in breeding applications, as has been done in other crops [35]. 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 maker 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 regards to achieving a specific morpho-type associated market class. This cost could be mitgated through the use of markers to recover morpho-type expediciously during pre-breeding. Ultimately, the judicious infusion of diversity into a breeding program is necessary for sustaining long-term gain.

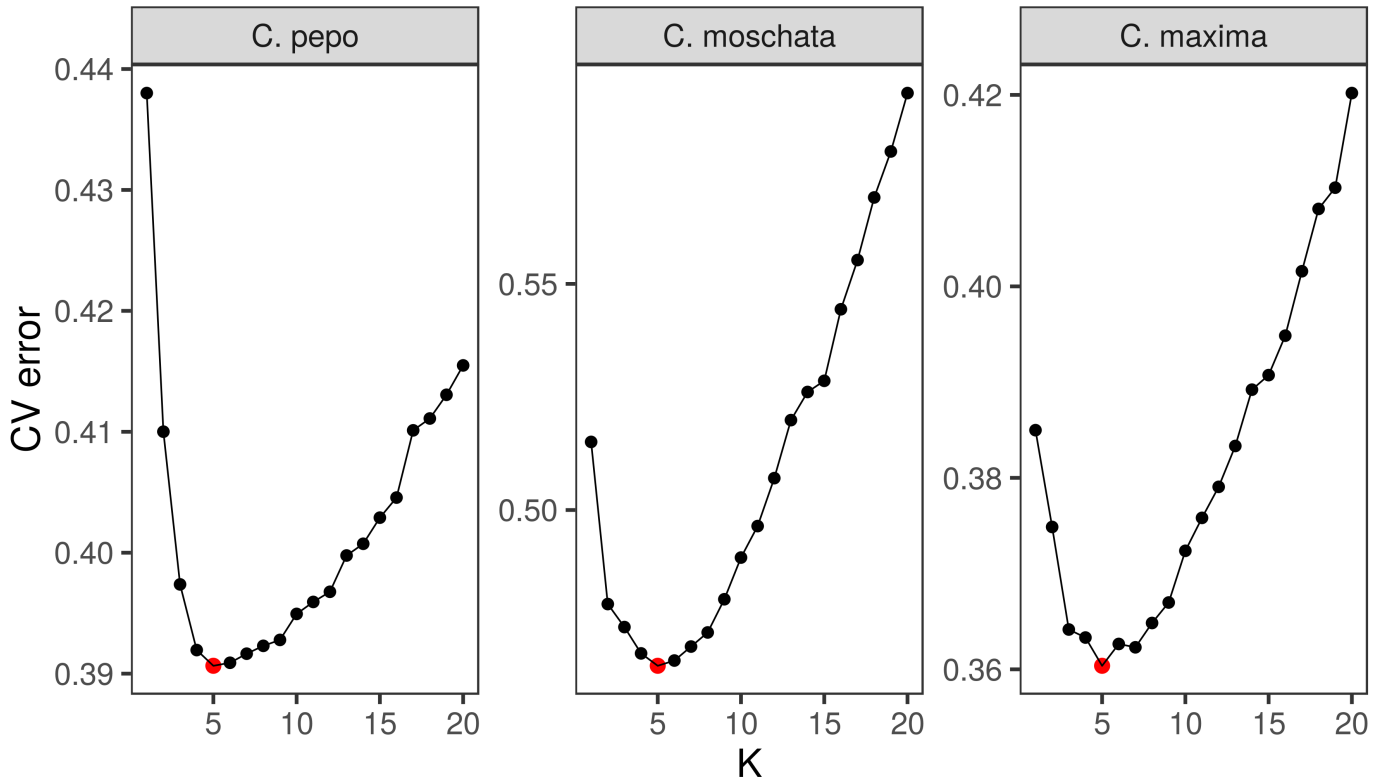
Genomic selection (GS) was proposed over twenty years ago [36], and has since become a standard breeding technique. Yet, to our knowledge, GS is not used to any appreciable degree by any of the public-sector breeding programs working with cucurbits. Studies specifically looking at GS in squash have demonstrated, as with every 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 [9,37,38]. Since cucurbit crops are more seed-limited than space-limited, a predict-part-test-part or sparse testing strategy is an obvious starting point [**doi?** 10.3389/fpls.2021.658978]. 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 idependently by interested breeding programs, rapidly convered into an amplicon-based assay, and tested in target germplasm. This set can then be used for routine genotyping, which is a necessary first step towards implementing GS [39].

At the interface of breeding and biology lies the phenomena of heterosis in squash. Although there is some evidence of heterosis in squash, the basis of this heterosis is not well understood. Unlike many other outcrossing monoicous crops such as, maize and onion, cultivars from Cucurbita, similar to sunflower, do not suffer from debilatating inbreeding. With little inbreeding depression, it would stand that lttle better-parent heterosis could be achieved considering the dominance theory of heterosis. Initial papers suggested that inbreeding in Cucurbita may not simply reduce yield as inbred

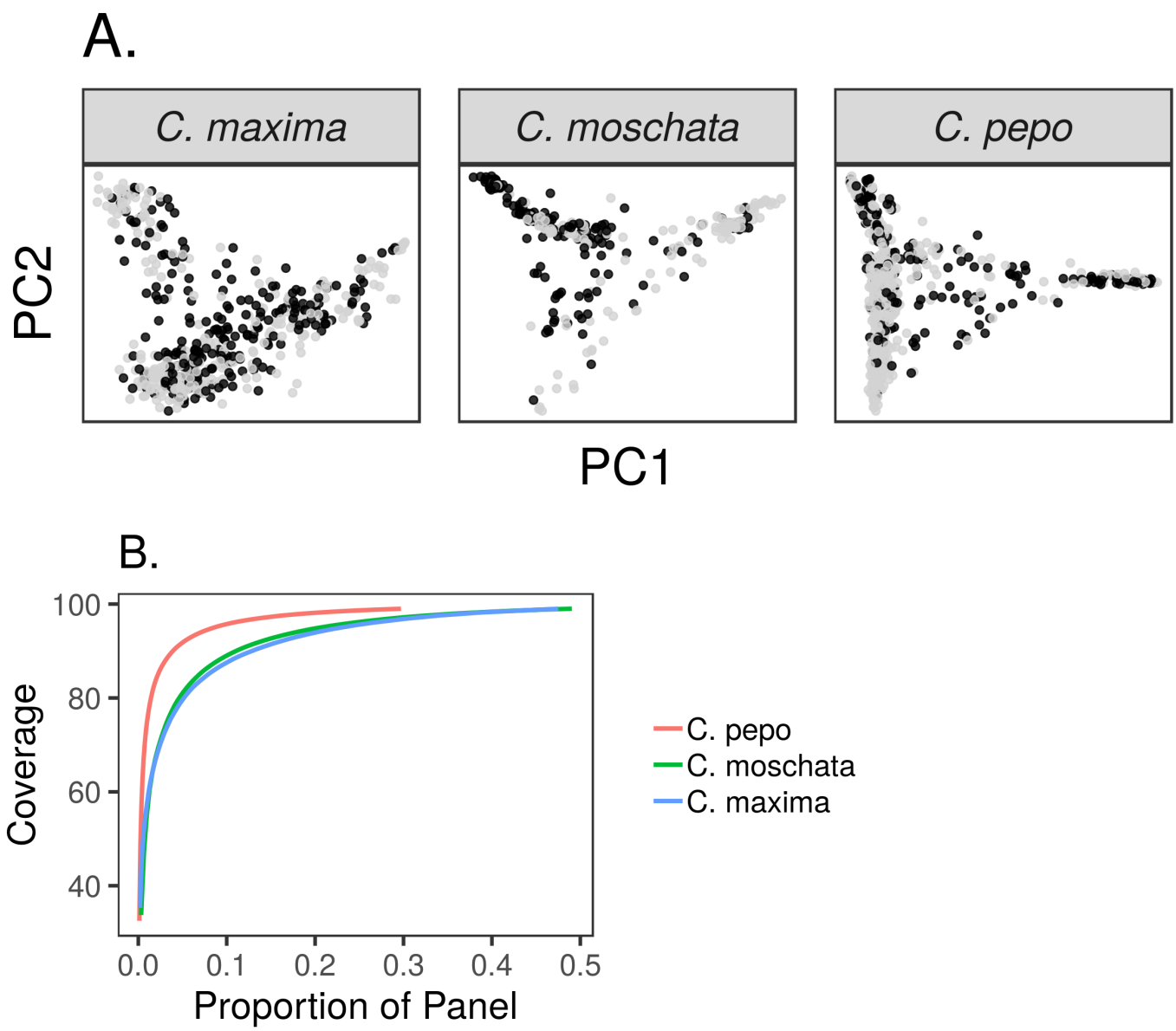
varieties have the capacity to compete with commercial check cultivars; however, better-parent heterosis has been observed in a Cucurbita pepo and C ucurbita maxima. Further, interspecific-heterosis has been observed at the gene-expression level in *C moschata x C. maxima* hybrids [4].

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 resistance and developing resistant breeding lines [40,41]. Further, if a trait segregates closely with population sructure, 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 [42]. When higher density genotyping may be necessary or the PIs are not well charaterized for a trait of interest, the data generated in this study can be used to prioritize accessions for re-sequencing and phentyping. Our geno-core analysis provides a subset of several hundred accessions that would likely be informative for re-sequencing efforts.

Supplemental Figures



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



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.

References

1. **Pumpkin and Winter Squash**
María Ferriol, Belén Picó
Springer Science and Business Media LLC (2007-12-06) https://doi.org/dmqkmf
DOI: 10.1007/978-0-387-30443-4_10

2. **The Genes of Pumpkin and Squash**
Harry S Paris, Rebecca Nelson Brown
HortScience (2005-10) https://doi.org/gmkkfh
DOI: 10.21273/hortsci.40.6.1620

3. **Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology**
Guillaume Chomicki, Hanno Schaefer, Susanne S Renner
New Phytologist (2019-08) https://doi.org/gsg7
DOI: 10.1111/nph.16015 · PMID: 31230355

4. **Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes**
Honghe Sun, Shan Wu, Guoyu Zhang, Chen Jiao, Shaogui Guo, Yi Ren, Jie Zhang, Haiying Zhang, Guoyi Gong, Zhangcai Jia, … Yong Xu
Molecular Plant (2017-10) https://doi.org/gb4cx2
DOI: 10.1016/j.molp.2017.09.003 · PMID: 28917590

5. **Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops**
Yi Zheng, Shan Wu, Yang Bai, Honghe Sun, Chen Jiao, Shaogui Guo, Kun Zhao, Jose Blanca, Zhonghua Zhang, Sanwen Huang, … Zhangjun Fei
Nucleic Acids Research (2019-01-08) https://doi.org/gmcmq9
DOI: 10.1093/nar/gky944 · PMID: 30321383 · PMCID: PMC6324010

6. **An SNP-based saturated genetic map and QTL analysis of fruit-related traits in Zucchini using Genotyping-by-sequencing**
Javier Montero-Pau, José Blanca, Cristina Esteras, Eva Ma Martínez-Pérez, Pedro Gómez, Antonio J Monforte, Joaquín Cañizares, Belén Picó
BMC Genomics (2017-01-18) https://doi.org/gmkkvf
DOI: 10.1186/s12864-016-3439-y · PMID: 28100189 · PMCID: PMC5241963

7. **A high-density linkage map and QTL mapping of fruit-related traits in pumpkin (Cucurbita moschata Duch.)**
Yu-Juan Zhong, Yang-Yang Zhou, Jun-Xing Li, Ting Yu, Ting-Quan Wu, Jian-Ning Luo, Shao-Bo Luo, He-Xun Huang
Scientific Reports (2017-10-06) https://doi.org/gmkktr
DOI: 10.1038/s41598-017-13216-3 · PMID: 28986571 · PMCID: PMC5630576

8. **Genetic mapping of ovary colour and quantitative trait loci for carotenoid content in the fruit of Cucurbita maxima Duchesne**
Karolina Kaźmińska, Ewelina Hallmann, Anna Rusaczonek, Aleksandra Korzeniewska, Mirosław Sobczak, Joanna Filipczak, Karol Seweryn Kuczerski, Jarosław Steciuk, Monika Sitarek-Andrzejczyk, Marek Gajewski, … Grzegorz Bartoszewski
Molecular Breeding (2018-08-27) https://doi.org/gd6tc4
DOI: 10.1007/s11032-018-0869-z · PMID: 30237748 · PMCID: PMC6133072

9. **Genomic Prediction of Pumpkin Hybrid Performance**
Po-Ya Wu, Chih-Wei Tung, Chieh-Ying Lee, Chen-Tuo Liao
The Plant Genome (2019-06) https://doi.org/gmkkvg
DOI: 10.3835/plantgenome2018.10.0082 · PMID: 31290920

10. **Whole-genome resequencing of Cucurbita pepo morphotypes to discover genomic variants associated with morphology and horticulturally valuable traits**
Aliki Xanthopoulou, Javier Montero-Pau, Ifigeneia Mellidou, Christos Kissoudis, José Blanca, Belén Picó, Aphrodite Tsaballa, Eleni Tsaliki, Athanasios Dalakouras, Harry S Paris, … Ioannis Ganopoulos
Horticulture Research (2019-08-11) https://doi.org/gmkkvd
DOI: 10.1038/s41438-019-0176-9 · PMID: 31645952 · PMCID: PMC6804688

11. **Genomic Prediction and Selection for Fruit Traits in Winter Squash**
Christopher O Hernandez, Lindsay E Wyatt, Michael R Mazourek
G3 Genes|Genomes|Genetics (2020-10-01) https://doi.org/gmkzhb
DOI: 10.1534/g3.120.401215 · PMID: 32816923 · PMCID: PMC7534422

12. **Mobilizing Crop Biodiversity**
Susan McCouch, Zahra Katy Navabi, Michael Abberton, Noelle L Anglin, Rosa Lia Barbieri, Michael Baum, Kirstin Bett, Helen Booker, Gerald L Brown, Glenn J Bryan, … Loren H Rieseberg
Molecular Plant (2020-10) https://doi.org/gmkzrd
DOI: 10.1016/j.molp.2020.08.011 · PMID: 32835887

13. **The USDA cucumber (Cucumis sativus L.) collection: genetic diversity, population structure, genome-wide association studies, and core collection development**
Xin Wang, Kan Bao, Umesh K Reddy, Yang Bai, Sue A Hammar, Chen Jiao, Todd C Wehner, Axel O Ramírez-Madera, Yiqun Weng, Rebecca Grumet, Zhangjun Fei
Horticulture Research (2018-10-01) https://doi.org/gfdjfd
DOI: 10.1038/s41438-018-0080-8 · PMID: 30302260 · PMCID: PMC6165849

14. **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**
Shan Wu, Xin Wang, Umesh Reddy, Honghe Sun, Kan Bao, Lei Gao, Linyong Mao, Takshay Patel, Carlos Ortiz, Venkata L Abburi, … Zhangjun Fei
Plant Biotechnology Journal (2019-05-07) https://doi.org/gmkktt
DOI: 10.1111/pbi.13136 · PMID: 31022325 · PMCID: PMC6835170

15. **Morpho-Physiological Aspects of Productivity and Quality in Squash and Pumpkins (Cucurbita spp.)**
JBrent Loy
Critical Reviews in Plant Sciences (2004-07) https://doi.org/abs/10.1080/07352680490490733
DOI: abs/10.1080/07352680490490733

16. **Germplasm enhancement of Cucurbita pepo (pumpkin, squash, gourd: Cucurbitaceae): progress and challenges**
Harry S Paris
Euphytica (2015-11-24) https://doi.org/f8ds6k
DOI: 10.1007/s10681-015-1605-y

17. **Evolutionary and domestication history of Cucurbita (pumpkin and squash) species inferred from 44 nuclear loci**
Heather R Kates, Pamela S Soltis, Douglas E Soltis
Molecular Phylogenetics and Evolution (2017-06) https://doi.org/f97dq2

DOI: 10.1016/j.ympev.2017.03.002 · PMID: 28288944

18. **Italian horticultural and culinary records of summer squash (Cucurbita pepo , Cucurbitaceae) and emergence of the zucchini in 19th-century Milan**
Teresa A Lust, Harry S Paris
Annals of Botany (2016-07) https://doi.org/gmkk6b
DOI: 10.1093/aob/mcw080 · PMID: 27343231 · PMCID: PMC4934399

19. **Parallel Evolution Under Domestication and Phenotypic Differentiation of the Cultivated Subspecies of Cucurbita pepo (Cucurbitaceae)**
Harry S Paris, Ales Lebeda, Eva Křistkova, Thomas C Andres, Michael H Nee
Economic Botany (2012-01-31) https://doi.org/fzc57g
DOI: 10.1007/s12231-012-9186-3

20. **The making of giant pumpkins: how selective breeding changed the phloem of C ucurbita maxima from source to sink**
JESSICA A SAVAGE, DUSTIN F HAINES, NMICHELE HOLBROOK
Plant, Cell & Environment (2015-08) https://doi.org/f7jhh7
DOI: 10.1111/pce.12502 · PMID: 25546629

21. **TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline**
Jeffrey C Glaubitz, Terry M Casstevens, Fei Lu, James Harriman, Robert J Elshire, Qi Sun, Edward S Buckler
PLoS ONE (2014-02-28) https://doi.org/f5zjsk
DOI: 10.1371/journal.pone.0090346 · PMID: 24587335 · PMCID: PMC3938676

22. **Fast and accurate short read alignment with Burrows-Wheeler transform**
H Li, R Durbin
Bioinformatics (2009-05-18) https://doi.org/dqt59j
DOI: 10.1093/bioinformatics/btp324 · PMID: 19451168 · PMCID: PMC2705234

23. **The variant call format and VCFtools**
P Danecek, A Auton, G Abecasis, CA Albers, E Banks, MA DePristo, RE Handsaker, G Lunter, GT Marth, ST Sherry, ... 1000 Genomes Project Analysis Group
Bioinformatics (2011-06-07) https://doi.org/b6kxfd
DOI: 10.1093/bioinformatics/btr330 · PMID: 21653522 · PMCID: PMC3137218

24. **Enhancements to the ADMIXTURE algorithm for individual ancestry estimation**
David H Alexander, Kenneth Lange
BMC Bioinformatics (2011-06-18) https://doi.org/dtnztg
DOI: 10.1186/1471-2105-12-246 · PMID: 21682921 · PMCID: PMC3146885

25. **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses**
Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW de Bakker, Mark J Daly, Pak C Sham
The American Journal of Human Genetics (2007-09) https://doi.org/cp2rzn
DOI: 10.1086/519795 · PMID: 17701901 · PMCID: PMC1950838

26. **Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer**
Giovanny Covarrubias-Pazaran
PLOS ONE (2016-06-06) https://doi.org/ggjp6v
DOI: 10.1371/journal.pone.0156744 · PMID: 27271781 · PMCID: PMC4894563

27. **SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data**
Tae-Ho Lee, Hui Guo, Xiyin Wang, Changsoo Kim, Andrew H Paterson
BMC Genomics (2014) https://doi.org/f5xgnj

DOI: 10.1186/1471-2164-15-162 · PMID: 24571581 · PMCID: PMC3945939

28. **Genomic Heritability: What Is It?**
Gustavo de los Campos, Daniel Sorensen, Daniel Gianola
PLOS Genetics (2015-05-05) https://doi.org/gmkw9r
DOI: 10.1371/journal.pgen.1005048 · PMID: 25942577 · PMCID: PMC4420472

29. **Genome-enabled predictions for binomial traits in sugar beet populations**
Filippo Biscarini, Piergiorgio Stevanato, Chiara Broccanello, Alessandra Stella, Massimo Saccomani
BMC Genetics (2014) https://doi.org/f6jhz3
DOI: 10.1186/1471-2156-15-87 · PMID: 25053450 · PMCID: PMC4113669

30. **Log-likelihood-based Pseudo- R ² in Logistic Regression**
Giselmar AJ Hemmert, Laura M Schons, Jan Wieseke, Heiko Schimmelpfennig
Sociological Methods & Research (2016-03-18) https://doi.org/gdzt5m
DOI: 10.1177/0049124116638107

31. **LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms**
Daniel Money, Kyle Gardner, Zoë Migicovsky, Heidi Schwaninger, Gan-Yuan Zhong, Sean Myles
G3 Genes|Genomes|Genetics (2015-11-01) https://doi.org/f79ns2
DOI: 10.1534/g3.115.021667 · PMID: 26377960 · PMCID: PMC4632058

32. **TASSEL: software for association mapping of complex traits in diverse samples**
PJ Bradbury, Z Zhang, DE Kroon, TM Casstevens, Y Ramdoss, ES Buckler
Bioinformatics (2007-06-22) https://doi.org/fdj9qg
DOI: 10.1093/bioinformatics/btm308 · PMID: 17586829

33. **GenoCore: A simple and fast algorithm for core subset selection from large genotype datasets**
Seongmun Jeong, Jae-Yoon Kim, Soon-Chun Jeong, Sung-Taeg Kang, Jung-Kyung Moon, Namshin Kim
PLOS ONE (2017-07-20) https://doi.org/gbn84s
DOI: 10.1371/journal.pone.0181420 · PMID: 28727806 · PMCID: PMC5519076

34. **A high-density genetic map for anchoring genome sequences and identifying QTLs associated with dwarf vine in pumpkin (Cucurbita maxima Duch.)**
Guoyu Zhang, Yi Ren, Honghe Sun, Shaogui Guo, Fan Zhang, Jie Zhang, Haiying Zhang, Zhangcai Jia, Zhangjun Fei, Yong Xu, Haizhen Li
BMC Genomics (2015-12-24) https://doi.org/gb3hqt
DOI: 10.1186/s12864-015-2312-8 · PMID: 26704908 · PMCID: PMC4690373

35. **1k-RiCA (1K-Rice Custom Amplicon) a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice**
Juan David Arbelaez, Maria Stefanie Dwiyanti, Erwin Tandayu, Krizzel Llantada, Annalhea Jarana, John Carlos Ignacio, John Damien Platten, Joshua Cobb, Jessica Elaine Rutkoski, Michael J Thomson, Tobias Kretzschmar
Rice (2019-07-26) https://doi.org/c8vh
DOI: 10.1186/s12284-019-0311-0 · PMID: 31350673 · PMCID: PMC6660535

36. **Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps**
THE Meuwissen, BJ Hayes, ME Goddard
Genetics (2001-04-01) https://doi.org/gknztd
DOI: 10.1093/genetics/157.4.1819

37. **Genomic Prediction and Selection for Fruit Traits in Winter Squash**

Christopher O Hernandez, Lindsay E Wyatt, Michael R Mazourek
G3 Genes|Genomes|Genetics (2020-10-01) https://doi.org/gmkzhb
DOI: doi.org/10.1534/g3.120.401215

38. **Evaluation of Selection Methods for Resistance to a Specialist Insect Pest of Squash (Cucurbita pepo)**
Lauren J Brzozowski, Michael Mazourek
Agronomy (2020-06-14) https://doi.org/gmndwq
DOI: doi.org/10.3390/agronomy10060847

39. **Strategies for Effective Use of Genomic Information in Crop Breeding Programs Serving Africa and South Asia**
Nicholas Santantonio, Sikiru Adeniyi Atanda, Yoseph Beyene, Rajeev K Varshney, Michael Olsen, Elizabeth Jones, Manish Roorkiwal, Manje Gowda, Chellapilla Bharadwaj, Pooran M Gaur, … Kelly R Robbins
Frontiers in Plant Science (2020-03-27) https://doi.org/gmfpfj
DOI: 10.3389/fpls.2020.00353 · PMID: 32292411 · PMCID: PMC7119190

40. **A combined BSA-Seq and linkage mapping approach identifies genomic regions associated with Phytophthora root and crown rot resistance in squash**
Gregory Vogel, Kyle E LaPlant, Michael Mazourek, Michael A Gore, Christine D Smart
Theoretical and Applied Genetics (2021-01-03) https://doi.org/gmndwj
DOI: 10.1007/s00122-020-03747-1 · PMID: 33388885

41. **Performance and Resistance to Phytophthora Crown and Root Rot in Squash Lines**
Kyle E LaPlant, Gregory Vogel, Ella Reeves, Christine D Smart, Michael Mazourek
HortTechnology (2020-10) https://doi.org/gmndwp
DOI: 10.21273/horttech04636-20

42. **Divergence of defensive cucurbitacins in independent domestication events leads to differences in specialist herbivore preference**
Lauren J Brzozowski, Michael A Gore, Anurag A Agrawal, Michael Mazourek
Plant, Cell & Environment (2020-09-12) https://doi.org/gmcbhx
DOI: 10.1111/pce.13844 · PMID: 32666553