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

This manuscript (permalink) was automatically generated from ch728/cucurbit-usda@1aa490f on August 25, 2021.

Authors

- **Christopher Owen Hernandez**

  Department of Plant Breeding and Genetics, Cornell University · Funded by Grant XXXXXXXX

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 *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]; 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 by geneticists for mapping studies. Many s. The collections of The Cucurbit Coordinated Agricultural Project (CucCap project) has been 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 [11,12].

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 [13]. 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 [14]. 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 [15].

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* [15]. Europe played a crucial role as a secondary center of diversification for *pepo*, but not *ovifera* [16]. 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 [17].

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 [18]. 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 [15]. *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 [19]. Tags produced by this pipeline were aligned using the default settings of the BWA aligner [20]. Raw variants were filtered using VCFtools [21]. 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 [22], 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 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). 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. The 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 . 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 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 , , and . Data included phenotype data as well as narratives/descriptions associated with accessions. Narrative data were parsed into a list of informative words, using a

custom Python script, to produce a qualitative snapshot of the diversity present in each collection. 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 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 (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 $h_g^2 = \frac{\sigma_g^2}{\sigma_g^2+\frac{\pi^2}{3}}$. 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 .

GWAS

Data were imputed prior to association analysis. LinkImpute , as implemented by the TASSEL "LDKNNiImputatioHetV2Plugin" plugin was used for imputation with default settings. Any data still missing after this process were mean imputed. The GENESIS 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.

Creation of a Core Collection

Subsets representative of each panel's genetic diversity were identified through running GenoCore on each of the filtered SNP sets. A subset of the panel and key genotypes from the other two species were combined to form a core collection for the cucurbit community. (Insert criteria here) . These genotypes will be further purified through two additional rounds of selfing and then resequenced using skim-sequecing to produce whole-genome data.

Results

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. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **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

20. **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

21. **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

22. **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