Characterization of Gene Presence/Absence Variations in Wild and Cultivated Watermelons

Cornell University

Benjamin Beer, Shan Wu, Zhangjun Fei

BTI

Introduction

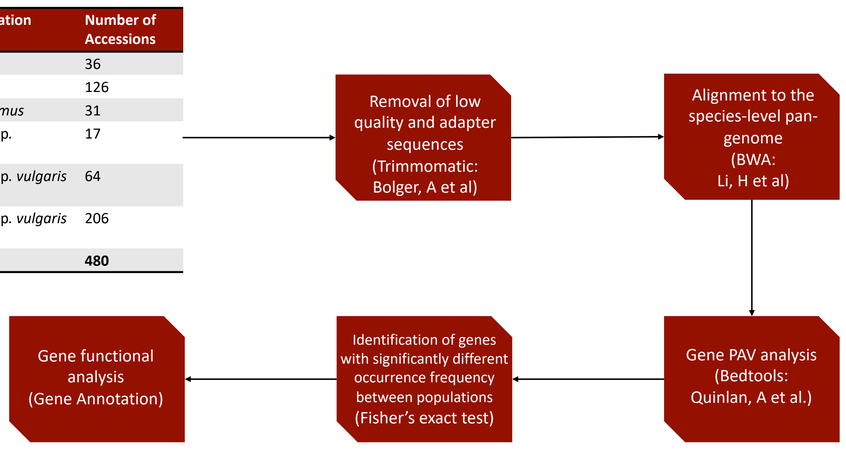
Watermelon *(Citrullus lanatus* subsp. *vulgaris)* is a prevalent crop for consumption in many countries worldwide. Since its domestication over 4,000 years ago it has found popularity in many countries, due in part to its nutritional value and appetizing taste (Maoto et al. 2018). Despite this long history of production growers still struggle to produce watermelon with negligible crop loss (Dube et al. 2020).

One of the leading causes for this crop loss is watermelon's high susceptibility to diseases. Some prevalent diseases include varieties of fungal diseases along with viral infections. However, wild species in the *Citrullus* genus such as *C. mucosospermus, C. amarus, and C. colocynthis* exhibit resistances to many of these diseases (Paris 2015). Introduction of specific genes from wild watermelons may produce the resistances needed to improve watermelon yield.

To successfully utilize the genetic diversity preserved in the wild watermelons and guide more efficient selection of breeding materials that carry beneficial traits, characterization of genes presence/absence variations (PAVs) in wild and cultivated watermelons is necessary. A pan-genome of watermelons has been constructed to capture genes existing in different watermelon species. Comparative analyses using this pan-genome can reveal PAVs of functionally important genes that may be selected or lost during watermelon domestication and those may confer disease resistance in the wild watermelons.

Materials & Methods

Species/Population	Number of Accessions
C. colocynthis	36
C. amarus	126
C. mucosospermus	31
C. lanatus subsp. *cordophanus*	17
C. lanatus subsp. *vulgaris* landrace	64
C. lanatus subsp. *vulgaris* cultivar	206
Total	**480**



Results

Characterization of the watermelon super pan-genome

Gene PAVs were first analyzed in the four species-level pan-genomes. In *C. lanatus*, 86.23% were shared among all accessions (core) while 3.17% were present in all but one (softcore); 10.18% of genes were present in at least two accessions (shell) and 0.32% were present in a single accession (private) (**Figure 1**). Other three species exhibited similar characteristics, with the core gene content higher than 85%. These high core gene content values suggested a relatively low gene PAV diversity among individuals within the same watermelon species.

Characterizing the genus-level *Citrullus* super pan-genome revealed a core gene content of 63.68%, and rest of the genes belonged to the accessory genome (**Figure 2**), demonstrating the divergence among watermelon species. When looking to find if a gene is present in a species, 75% of genes could be found in all species while 8% were unique to a single given species, and 16% of genes were shared between all but one species (**Figure 3**). This displayed diversity should offer opportunities for the introduction of beneficial and disease resistant genes from wild species to the cultivated watermelon.

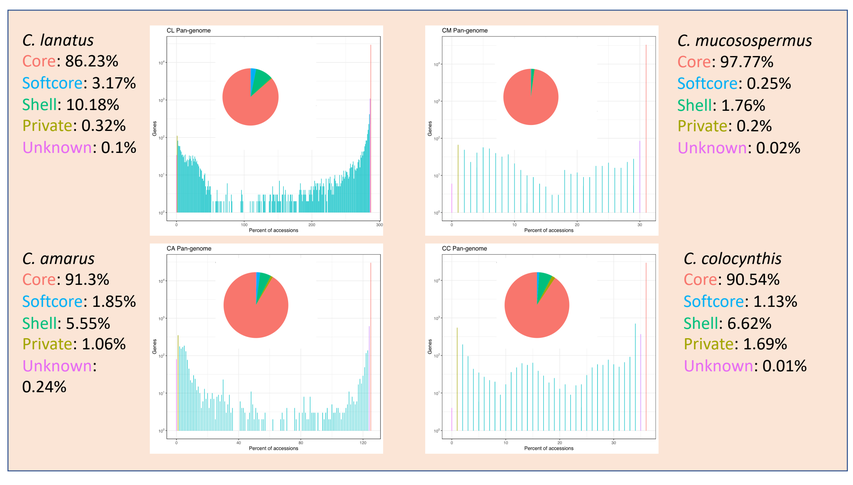
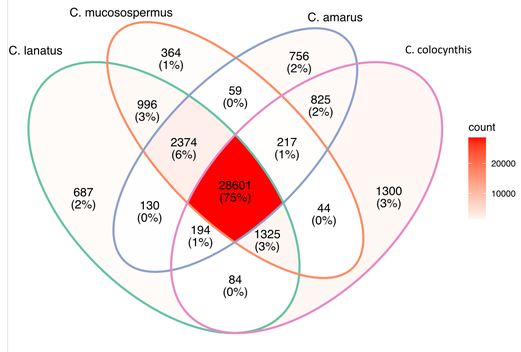


Figure 1 (Top). Histograms and pie charts showing features of the four species-specific pan-genomes

Figure 2 (Bottom Left). Histogram and pie chart showing features of the *Citrullus* super pan-genome

Figure 3 (Bottom Right). Venn diagram showing the numbers of genes present within the four watermelon species



Genes found in at least one accession of a species were considered as present in the species.

Gene with different occurrence frequencies between populations

To identify genes lost in *C. lanatus* (including the cultivated watermelon) but present and its disease resistant wild relatives, gene occurrence frequencies were compared between species. The significance of changed frequency was determined using Fisher's exact test with a P value less than 0.001 and a fold change of more than two (**Figure 4a-c**; **Table 1**). *C. amarus* and *C. colocynthis* seemed to have the largest divergence in gene content compared to *C. lanatus*. Namely, *C. amarus* had the most genes with significantly different occurrence frequencies: 5,365 (**Table 1**).

Genes with changed frequencies during domestication were also identified by comparing the cultivated watermelon (*C. lanatus* subsp. *vulgaris*) to its wild progenitor (*C. lanatus* subsp. *cordophanus*) (**Figure 4d**; **Table 1**). Very few genes had changed frequencies during improvement (from landrace to cultivars) (**Figure 4e**; **Table 1**).

Table 1. Number of genes with significantly change frequencies.

		Number of genes with significantly changed frequency in population 2	
Population 1	**Population 2**	**Increased**	**Decreased**
C. colocynthis	*C. lanatus*	2295	1293
C. amarus	*C. lanatus*	3470	1895
C. mecosospermus	*C. lanatus*	251	1011
C. lanatus co.	*C. lanatus lr.*	14	130
C. lanatus lr.	*C. lanatus cu.*	3	32

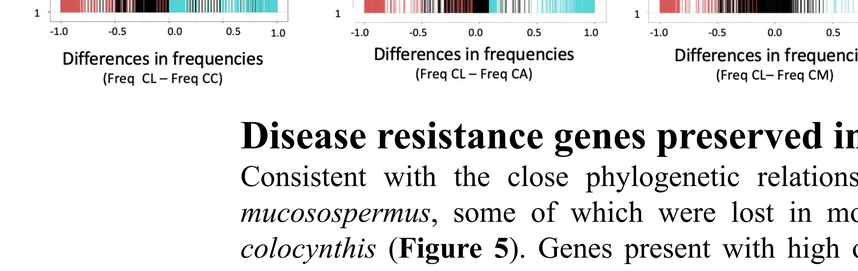
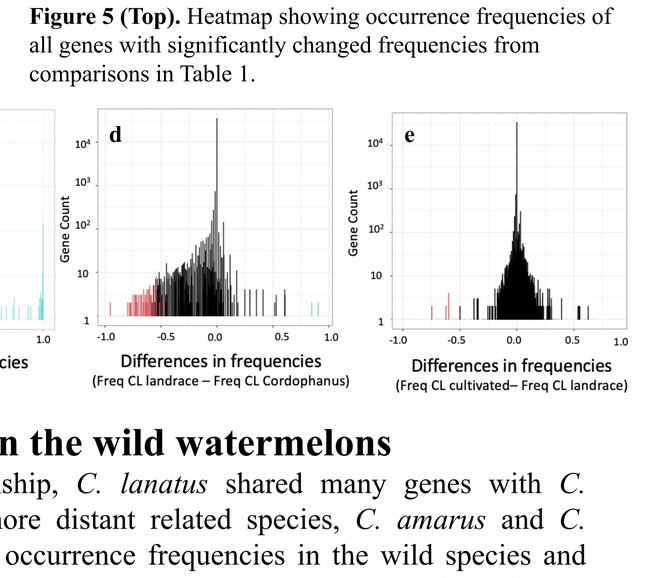


Figure 4 (Bottom). Distribution of differences in occurrence frequencies. Colored bars represent genes with significantly altered frequencies between populations.



CLcu, *C. lanatus* subsp. *vulgaris* cultivar; CLlr, *C.lanatus* subsp. *vulgaris* landrace; CLco, *C. lanatus* subsp. *cordophanus;* CM, *C. mucosospermus;* CA, *C. amarus;* CC, *C. colocynthis*

Figure 5 (Top). Heatmap showing occurrence frequencies of all genes with significantly changed frequencies from comparisons in Table 1.

Disease resistance genes preserved in the wild watermelons

Consistent with the close phylogenetic relationship, *C. lanatus* shared many genes with *C. mucosospermus*, some of which were lost in more distant related species, *C. amarus* and *C. colocynthis* (**Figure 5**). Genes present with high occurrence frequencies in the wild species and mostly lost in the cultivated watermelons were also observed (**Figure 5**). A total of 24 disease resistant genes were almost exclusively found in the wild watermelons especially in *C. amarus* and *C. colocynthis* (**Figure 6**). Two disease resistant genes were found to exist at a higher frequency in the direct progenitor, *C. lanatus* subsp. *cordophanus* and were completely absent form *C. lanatus* subsp. *vulgaris* (**Table 2**).

Figure 6. Heatmap showing normalized occurrence frequencies of disease resistance genes with changed occurrence frequencies among populations

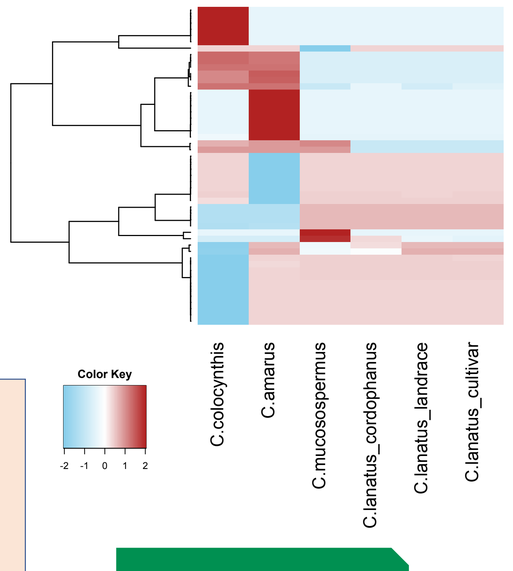


Table 2. Disease resistance genes present prevalently in the wild watermelons.

Name	Function	C. colocynthis	C. amarus	C. mucosospermus	C.lanatus_subsp.cordophanus	C. lanatus_landrace	C. lanatus_cultivar
GrpOrp_036826	pathogen-related protein-like	0	0.5476	0	0	0	0
GrpOrp_037226	Pathogenesis-related thaumatin family protein	1	0	0	0	0	0
GrpOrp_037288	pathogenesis-related protein-1-like	1	0	0	0	0	0
GrpOrp_037368	pathogenesis-related protein 1-like	1	0	0	0	0	0
GrpOrp_038229	basic form of pathogenesis-related protein 1-like	0.7778	0	0	0	0	0
GrpOrp_039023	basic form of pathogenesis-related protein 1-like	0	0	1	0	0	0
GrpSep_035585	pathogenesis-related protein-1-like	0	1	0	0	0	0
GrpSep_035586	basic form of pathogenesis-related protein 1-like	0	1	0	0	0	0
GrpSep_035591	pathogenesis-related protein-1-like	0	1	0	0	0	0
GrpSep_036414	basic form of pathogenesis-related protein 1-like	0	0.9841	0	0	0	0
GrpSyn_028171	Pathogenesis-related protein-1-like protein	1	0.9921	0	0	0	0
GrpSyn_032290	Pathogenesis-related protein-1-like protein	1	0.9365	0	0	0	0
GrpOrp_037567	Resistance gene-like protein	1	0	0	0	0	0
GrpOrp_037728	NB-ARC domain-containing disease resistance protein	1	0	0	0	0	0
GrpOrp_036727	TMV resistance protein N-like	0	0.7222	0	0	0	0
GrpOrp_036728	Resistance gene-like protein	0	0.7222	0	0	0	0
GrpSyn_032294	Resistance gene-like protein	0.0556	0.7302	0	0	0	0
GrpSyn_030406	Resistance gene-like protein	0.8056	0.9127	1	0	0	0
GrpSyn_030271	LEAF RUST 10 DISEASE-RESISTANCE LOCUS RECEPTOR-LIKE PROTEIN KINASE-like 2.4	1	1	1	0	0	0
GrpSyn_032957	Resistance gene-like protein	1	0.9365	0	0	0	0
GrpSyn_032283	TMV resistance protein N	0.7778	0.9762	0	0	0	0
GrpSyn_032862	TIR-NBS-LRR disease resistance protein	0.8056	0.9762	0	0	0	0
GrpSep_035078	disease resistance RPP13-like protein 4	0	0	1	0.4118	0.1406	0.0971
GrpSyn_031172	pathogen-related protein-like	1	1	0	0.2353	0.0781	0.1845

Summary

In this study, we characterized PAVs of genes in the watermelon super pan-genome, which demonstrated the divergence among the wild and cultivated watermelons. Genes with significantly different occurrence frequencies among the different species and populations were identified. These genes included disease resistance genes that were lost in the cultivated watermelon, which could be brought back from the wild watermelons.

References

Bolger, A. et al. Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics* **30**, 2114-2120 (2014).

Dube, G. et al. Watermelon production in Africa: challenges and opportunities, *International Journal of Vegetable Science* **27**, 211-219 (2021).

Li, H. et al. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

Maoto M. et al. Watermelon as a potential fruit snack, *International Journal of Food Properties* **22**, 355-370 (2019).

Paris, H. Origin and emergence of the sweet dessert watermelon, *Citrullus lanatus*, *Annals of Botany* **116**, 133–148 (2015).

Quinlan, A. et al. BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics* **26**, 841–842 (2010).

Acknowledgments

This 2022 High School Research Internship Experience was made possible by Cornell University, USDA, NSF, and Boyce Thompson Institute as well as individual donors.

USDA NSF