MOLECULAR EVALUATION OF THE GENETIC DIVERSITY OF PAPAYA (*CARICA PAPAYA*) IN PUERTO RICO

by

Dianiris Luciano Rosario

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

MASTER OF SCIENCE
In
BIOLOGY

UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2017

Approved by:

_____	_____
Dimuth Siritunga, Ph.D. President, Graduate Committee	Date
_____	_____
Hugo E. Cuevas, Ph.D. Member, Graduate Committee	Date
_____	_____
Timothy Porch, Ph.D. Member, Graduate Committee	Date
_____	_____
Carlos Rodriguez Minguela, Ph.D. Member, Graduate Committee	Date
_____	_____
Luis O. Del Rio, Ph.D. Representative of Graduate Studies	Date
_____	
Matias Cafaro, Ph.D. Chairperson of the Department	_____ Date

Abstract

Native to Central America, papaya (*Carica papaya*) is one of the most cultivated fruit crops in tropical areas of the world. The commercial success of papaya is not only due to its high nutritional qualities but also due of its short generation time. Assessing the genetic diversity of papaya is an important aspect of conservation of this important plant genetic resources. However, knowledge on the genetic diversity of papaya in Puerto Rico is poorly understood. Therefore, 139 papaya accessions collected from all over Puerto Rico were evaluated using 23 Simple Sequence Repeat (SSR) markers and compared to 13 varieties from the USDA repository and 10 commercial varieties that served as references. A total of 214 alleles were identified having a mean observed heterozygosity (H_o) of 0.219. The Inbreeding coefficient (F) yielded a value of 0.565 and when evaluating the population structure of these accessions, 2 groups (k=2) were identified. An Unweighted Pair Group Method with Arithmetic Mean (UPGMA) dendogram showed no geographical organization within the unknown Puerto Rico samples. Moreover, Single Nucleotide Polymorphisms (SNPs) identification using Genotyping by Sequencing was also used to assess the genetic diversity of the same samples. We found a total of 4, 245 SNPs. A mean observed heterozygosity (H_o) of 0.226 and Inbreeding Coefficient (G_{is}) of 0.067 was recorded. In agreement with the SSR analyses, the population structure showed that the samples grouped in 2 clusters (k=2). Overall, this study contributes to the knowledge of papaya genetic diversity in the Caribbean region which will be useful for the conservation of papaya genetic resources.

Resumen

La papaya (*Carica papaya*), nativa de centro América, es uno de los cultivos frutales mas sembrados en las zonas tropicales del mundo. El éxito comercial de papaya no solo se debe a su alto valor nutritivo pero también recae en su corto tiempo de generación. Los análisis de diversidad genética componen un aspecto importante en la conservación de recursos fitogenéticos. Existe poco conocimiento acerca de la diversidad genética de papaya en Puerto Rico. Es por esto que se evaluaron 162 accesiones de papaya utilizando 23 microsatelites. De estas 162 accesiones, 139 son muestras desconocidas de Puerto Rico, 13 muestras del repositorio de USDA y 10 variedades comerciales. Se identificó un total de 214 alelos y una heterocigosidad observada promedio (H_o) de 0.219. El coeficiente de endogamia (F) mostró un valor de 0.565 y al evaluar la estructura de la población de estas accesiones, se identificaron 2 grupos (k=2). Un dendograma utilizando "Unweighted Pair Group Method with Arithmetic Mean" (UPGMA), no mostró organización geográfica entre las muestras desconocidas de Puerto Rico. Otro método que se utilizó para evaluar la diversidad genética de papaya en Puerto Rico, fue la identificación de polimorfismos de nucleótido simple (SNPs) utilizando "Genotyping by Sequencing". Encontramos un total de 4, 245 SNPs. La heterocigosidad observada (H_o) promedio fue 0.226 y el coeficiente de endogamia (G_{is}) 0.067. La estructura de la población de muestras resultó en 2 grupos. Este estudio contribuye al conocimiento de la diversidad genética de papaya en el Caribe y puede ser útil para la conservación de recursos fitogenéticos.

Dedication

To my parents Diana and Edgar. Thank you for your love and unconditional support along life and for always encouraging me to be a better person every day.

Acknowledgements

The completion of this project was possible due to a collective effort and support of many individuals.

First, I would like to thank my advisor Dr. Dimuth Siritunga for his mentorship and constant support along my undergraduate and graduate academic life. Thank you for giving me my first opportunity, for guiding me and providing me with the best tools to make me grow. For the exposure to research and everything that comes with it. Also for encouraging me to always do more and believing in me unconditionally.

To my graduate committee Dr. Timothy Porch, Dr. Carlos Rodriguez Minguela, and Dr. Hugo Cuevas. Thank you for always having open doors for guidance and advice.

Many thanks to Dr. Carlos Acevedo Suarez for being both a mentor and friend; for teaching me how to teach.

To my research mentor Lorraine Rodriguez Bonilla for not only teaching me the technical aspect of research but also offering me her friendship and unconditional support. Thank You!

To my labmates and friends for being there in the good and challenging times Cristina, Clara, Kevin, Edlin, Alejandro, Noraliz, Lumariz. Thank You Luis, for being the best undergraduate student and friend.

To my friends Mara and Luis Alexis for being the best support system and friends a person could have while pursuing a masters. Also to Margarette, Ruben, Lizbeth, Patricia, Ricardo, Amelia, Jose, Maraliz, and Jeysika; thank you for the great memories!

Table of Contents

List of Tables

List of Figures

Glossary of Terms

Abbreviation	Term
AFLP	Amplified Fragment Length Polymorphism
CAPS	Cleaved Amplified Polymorphic Sequences
CTAB	Cetyl trimethylammonium bromide
ddH2O	Deionized distilled water
DNA	Deoxyribonucleic Acid
FAO	Food and Agriculture Organization of the United Nations
GBS	Genotyping by Sequencing
GWAS	Genome Wide Association Studies
IBS	Identity by Similarity
ISSR	Inter Simple Sequence Repeats
LD	Linkage Disequilibrium
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PIC	Polymorphic Index Content
PRSV	Papaya ringspot virus
RADP	Random Amplification of Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
SCAR	Sequence-characterized amplified region
SNPs	Single Nucleotide Polymorphisms
SSCP	Single-strand conformation Polymorphism
SSR	Simple Sequence Repeats
STR	Short Tandem Repeats
TEMED	Tetramethylethylenediamine
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
USDA	United States Department of Agriculture
WGS	Whole-genome shotgun sequencing

Justification

Papaya is a tropical fruit crop used as a food commodity in numerous countries. Its origins are traced back to Central America, specifically to the south of Mexico and Nicaragua. This fruit offers many benefits such as a fast growth, high nutritional content (especially Vitamin A and C), and high production of latex which is commercially exploited to extract the proteolytic enzyme papain (Becker, 1958). In 2013, a total of 12,420,584.69 tons of papaya were produced worldwide. In the Caribbean, Puerto Rico occupies the third position of papaya production, with an estimated 8,852 tons being produced in 2013 (FAO). Even though Puerto Rico imports 80% of the food that is consumed, an exponential growth in agriculture gross production value has been shown through the last decade (FAO). Today papaya is the 29th most profitable crop, it is widely cultivated by the citizens (Departamento de Agricultura, 2009). Despite the popularity of this crop in Puerto Rico, the commercial production does not meet the local demand, leading to importation from the Dominican Republic (Morton, 1987).

In Puerto Rico, research on papaya has been limited to evaluating the yield and quality of some varieties and no studies have attempted to characterize the genetic diversity in the island (Goenaga *et al.*, 2001). Genetic diversity studies are important for the development of conservation and breeding programs as well as for preserving a sustainable agriculture system. Recently, there has been an emerging sustainable agriculture movement in Puerto Rico (Carro-Figueroa, 2014). In order to have a sustainable development of papaya, conservation and evaluation of the Caribbean germplasm is necessary mainly because the Caribbean islands constitute a zone of secondary diversification (OCampo, 2006). Although the genetic diversity of papaya population from the Caribbean region has been previously characterized (OCampo, 2006), Puerto Rican papaya population have not been evaluated yet. Therefore, herein it is proposed to investigate the genetic diversity of papaya in Puerto Rico using a molecular approaches based on microsatellite markers and Single Nucleotide Polymorphisms (SNPs).

CHAPTER 1 Introduction

Origins and Taxonomy

Carica papaya, or commonly known as papaya or paw paw, is a tropical fruit crop belonging to the *Caricaceae* family which includes 35 latex-containing species divided in six genera (Silva *et al.*, 2007; Badillo, 1971; 1993; 2000). Within *Caricaceae* the genera are *Carica*, *Jarilla*, *Horovitzia*, *Jacaratia*, *Vasconcellea*, and *Cylicomorpha* (da Silva, 2007; Badillo, 2002). All genera are thought to be originated in America with the exception of *Cylicomorpha* which is endemic of Africa. *Carica papaya* is the only species found in the genus *Carica* and is naturalized across the neotropics (Badillo, 1971; Carvalho, 2014). Papaya is thought to originate from south Mexico and/or Central America although no direct archeological evidence has been found (Carvalho and Renner, 2014). Indirect evidence such as analyzing the place of provenance of herbarium specimens around the world indicate that more than 50% of the samples were collected from Mexico and Central America specifically Nicaragua (Fuentes and Santamaria, 2014). Also the discovery of wild populations of papaya in isolated areas of the Yucatan peninsula and morphological comparison to the commercial variety Maradol, suggests similarity among the wild populations placing Maradol as an outgroup (Fuentes and Santamaria, 2014). A wild relative of papaya *V. cundinamarcensis* has been reported to be grown and naturalized in the Puerto Rican highlands (Morton, 1987).

Papaya Biology

Papaya is a semi-woody, latex producing, rapid growing perennial tropical giant herb (Jimenez *et al.,* 2002; Moore, 2014; da Silva, 2007). Its fruit is characterized by an oval to round, slightly pyriform shape that may weigh up to 9 kg and have a waxy thin skin (Morton, 1987). Its seed to seed generation ranges between 9 to 15 months. Fruit size, number, and germination rates are among the most important traits for commercial production of this crop (Fuentes, unpublished; Moore, 2014). The individual's sex is one of the key determinants for fruit size in papaya. Domesticated papaya, as we know it today, can be either a dioecious fruit crop with two possible sex forms: female or male or a gynodiceous fruit crop with three possible sex forms: male, female, and hermaphrodite (Figure 1). In commercial plantations it is preferred to have female and hermaphrodite (predominantly found) individuals due to the fruit shape and size (Carvalho *et al.,* 2012; VanBuren *et al.,* 2015). Interestingly, wild papaya populations are strictly dioecious (Carvalho and Renner, 2012). A recent study suggests that papaya domestication, specifically the ability of papaya to produce hermaphrodite individuals, resulted from the Mayans or other indigenous cultures ~4000 years ago (VanBuren, 2015). Sex determination of papaya is attributed to its unique

characteristic of having two types of sex chromosomes: Y for a male and Y^h for a hermaphrodite (Weintgartner, 2012). For this reason, papaya is used as a model system to study sex determination in plants (Aryal, 2014). It is known that self-fertilization, common in many hermaphrodite plants, leads to a fast loss of heterozygosity (Hamilton, 2009).

Papaya fruits are very rich in Vitamin A and Vitamin C and is considered one of the most nutritious among the 35 commonly consumed fruits in the United States of America (Ming *et al.*, 2007). It is one of the fruits targeted to battle Vitamin A deficiencies in developing countries. Its low caloric value but high nutrition makes it a good source of minerals such as magnesium, potassium, boron and copper (Hardisson *et al.,* 2001; Wall and Tripathi, 2014). Table 1 shows the nutritional value for 100g of edible fresh weigh of papaya fruit (Wall and Tripathi, 2014).

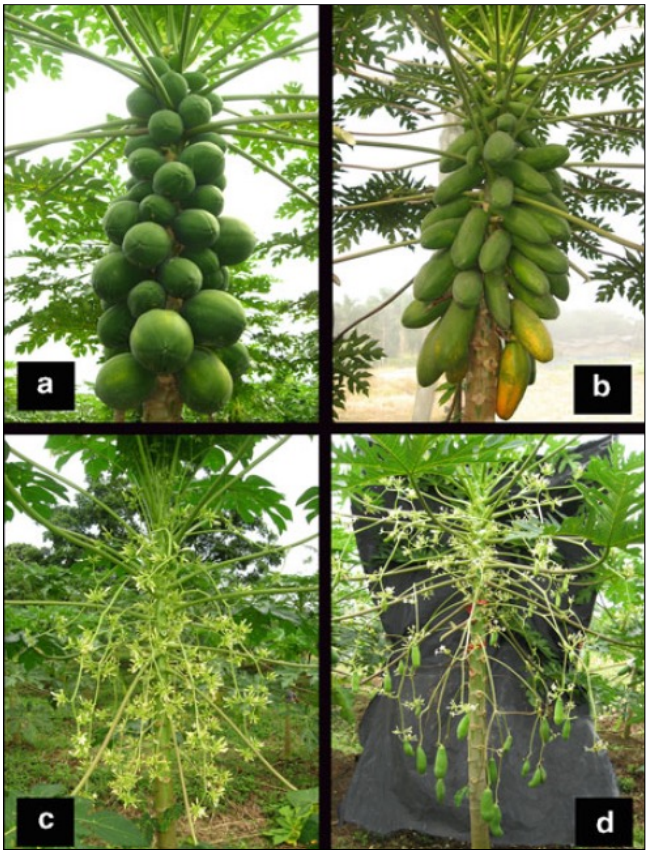


Figure 1 Papaya plants and their sex forms. *a) female; b) hermaphrodite; c) male; d) male fruit bearing. Source: Jiménez et al., 2013*

Table 1.Nutritional value of 100 grams of edible papaya (Wall and Tripathi, 2014)

Nutrient	Unit	Value per 100 g
Water	g	88.06
Energy	kcal	43
Protein	g	0.47
Total lipid (fat)	g	0.26
Carbohydrate, by difference	g	10.82
Fiber, total dietary	g	1.7
Sugars, total	g	7.82
Minerals		
Calcium, Ca	mg	20
Iron, Fe	mg	0.25
Magnesium, Mg	mg	21
Phosphorus, P	mg	10
Potassium, K	mg	182
Sodium, Na	mg	8
Zinc, Zn	mg	0.08
Vitamins		
Vitamin C, total ascorbic acid	mg	60.9
Thiamin	mg	0.023
Riboflavin	mg	0.027
Niacin	mg	0.357
Vitamin B-6	mg	0.038
Folate, DFE	µg	37
Vitamin B-12	µg	0
Vitamin A, RAE	µg	47
Vitamin A, IU	IU	950
Vitamin E (alpha-tocopherol)	mg	0.3
Vitamin D (D2 + D3)	µg	0
Vitamin D	IU	0
Vitamin K (phylloquinone)	µg	2.6
Lipids		
Fatty acids, total saturated	g	0.081
Fatty acids, total monounsaturated	g	0.072
Fatty acids, total polyunsaturated	g	0.058
Cholesterol	mg	0
Other		
Caffeine	Mg	0

Papaya Diseases: papaya ringspot virus

Many diseases arise from natural pathogens of papaya such as bacteria, nematode, virus, fungi and various insects (Kumar *et al.,* 2014). Although these pathogens cause biotic stress in papaya, the papaya ringspot virus (PRSV) is the most detrimental and has been researched heavily due to its well-documented negative effect on the Hawaiian production and economy (Gonsalves, 1998). Papaya ringspot virus belongs to the *Potyvirus* genus which is one of the largest and most important groups agriculturally, economically, and biologically (Ivanov *et al.,* 2014). PRSV has two types, one that infects papaya and cucurbits (PRSV-P) and another that infects cucurbits but not papaya (PRSV-W) (Bateson *et al.,* 2002). The infections are aphid-mediated in a non-persistent manner. Upon PRSV infection, papaya symptoms include mosaic and chlorosis in the leaves, ringspots in the fruit flesh and in severe cases leaf and fruit distortion (Gonsalves, 1998).

Domesticated papaya offers no resistance to PRSV although resistance genes have been identified in other species of the *Caricaceae* family such as the *Vasconcellea* genera. In order to confer resistance to PRSV in papaya, breeding of resistance genes from *Vasconcellea quercifolia* have been attempted with success limited to the identification of one resistant line (Siar *et al.,* 2010). The most successful and effective method to confer resistance to PRSV so far has been the development of two transgenic papaya varieties named 'Sun Up' and 'Rainbow' (Gonsalves, 1998). The genome of the 'Sun Up' variety has been sequenced (Ming *et al.,* 2008).

Papaya Production

In 2014, a total of 12,671,038 tons of papaya were produced worldwide with India being the leading producer (5,639,300) (Table 2). In the Caribbean, papaya is the 12th most produced commodity crop (864,995 tons), with the Dominican Republic being the major producer with 748,511.2 tons. Puerto Rico is the 4th largest producer in the region with an estimated production of 9,000 tons in 2014 (FAOSTAT, 2017).

Agriculture in Puerto Rico

The history of agriculture in Puerto Rico has embraced dramatic changes during the last century. During the early 1900s, commercial production of commodities like sugarcane, tobacco and cotton powered the local economy. Although the lands designated to plant food crops were decreasing as the century progressed, a significant amount of food was produced in the island. As a political consequence,

in the 1950s the main island economy shifted to a more industrialized and service-based economy rather than agriculture which eventually (in the 1990s) resulted in less production and more importation of food. Still, Puerto Rican government bets on the possibility of food imports ignoring the lack of control over food importation costs (Carro-Figueroa, 2014). Instead of strengthening and modernizing local supply chains, Puerto Rican supermarkets have generally sourced from low-cost global exporters and from the U.S. mainland (Setrini, 2012). Although in Puerto Rico, agriculture has not been a priority for the government during the last few decades, a sustainable agriculture movement has gradually emerged (Carro-Figueroa and Guptil, 1999). For sustainable agriculture to succeed in Puerto Rico it is vital to have diverse accessions of the important food crops that have the potential to improve the Puerto Rican economy. Thus, assessing and documenting the existing genetic diversity of food crops in the island is of great importance.

Table 2.Top 10 papaya producing countries and production value in tones. Source: FAO 2014

Country	Tons Produced
India	5,639,300
Brazil	1,603,351
Nigeria	850,000
Indonesia	840,121
Mexico	836,370
Dominican Republic	704,786
Democratic Republic of the Congo	220,483
Philippines	172,628
Venezuela	165,102
Thailand	157,571

Molecular Tools for Genetic Diversity Assessment

Genetic assessment of food crops has been historically done using morphological traits. Within the past 30 years, molecular techniques have been utilized to assess diversity due to its ability to evaluate DNA directly. In the more recent past, automation and cost efficiency of molecular tools have made the genetic diversity assessment more vigorous using techniques such as restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), amplified fragment length

polymoprhism (AFLP), single-nucleotide substitutions (SNPS) and simple sequence repeats (SSR) makers (Table 3; Argawal *et al*., 2008; Kesawat and Das, 2009; Schlotterer, 2004). Recently, with the emergence of Next Generation Sequencing (NGS) technologies, several methods such as genotyping by sequencing have been developed in order to assess the genetic diversity of populations.

Microsatellite Markers

Genetic molecular markers, are variants of DNA sequences at the same locus in the genomes of of two individuals that follow a Mendelian inheritance pattern (Kesawat and Das, 2009). The use of DNA-based molecular markers presents a more advantageous tool when compared to phenotypic markers due to the objective analysis that results from the data collection (Kesawat and Das, 2009). Researchers have established several criterias for the ideal molecular marker such as: (1) a polymorphic nature and even distribution along the genome; (2) a resolution capable of uncovering genetic differences; (3) the capacity to generate multiple independent markers; 4) simplicity, speed and low costs; (5) requirement of a small DNA amount and (6) co-dominant inheritance (Agarwal *et al.* 2008; Idrees and Irshad, 2014). Upon the development of different DNA-based molecular markers, one may choose the more suitable for answering the intended research question.).

Table 3.Different characteristics of the most commonly used molecular markers Source: Kesawat and Das (2009)

Characteristics	**RFLP**	**Mini Satellites**	**RAPD**	**Micro Satellites**	**ISSR**	**SSCP**	**CAPS**	**SCAR**	**AFLP**
Genomic Abundance	High	Medium	High	High	Medium-High	Low	Low	Low	High
Polymorphism Level	Medium	High	Medium	High	Medium	Low	Low-Medium	Medium	Medium
Locus Specificity	Yes	No/Yes	No	Yes	No	Yes	Yes	Yes	No
Co-dominance of alleles	Yes	No/Yes	Yes	No	No	Yes	Yes	No-Yes	No-Yes
Reproducibility	High	High	Low	High	Medium-High	Medium	High	High	Medium-High
Labor-Intensity	High	High	Low	Low	Low	Low-Medium	Low-Medium	Low	Medium
Technical Demands	High	High	Low	Low-Medium	Low-Medium	Medium	Low	Low	Medium
Operational Costs	High	High	Low	Low	Low-Medium	Low-Medium	Low	Low	Medium
Development Costs	Medium-High	Medium-High	Low-Medium	High	Low	High	Medium	Medium	Low
Required DNA Quantity	High	High	Low	Low	Low	Low	Low	Low	Medium
Amenability to automation	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes

Discovered during the early 1980s, microsatellites, also known as short tandem repeats (STR) or simple sequence repeats (SSR), are common nucleotide sequences repeated along both prokaryotic and eukaryotic genomes in both coding and non-coding regions. These repeats can vary from 1 to 6 nucleotides (Jarne 1996; Hoshino *et al*., 2012; Oliveira *et al*., 2006). Microsatellites can be found as mono-, di-, tri-, tetra-, penta- or hexa-nucleotide repeats (Schlotterer and Harr, 2001). Although it was thought that microsatellites were randomly distributed along genomes, recent studies suggest these molecular markers have a tendency to cluster. Also the frequency of microsatellites is inversely related to the repeat number having then a major number of smaller microsatellites distributed along the genome (Schlotterer and Harr, 2001). Their discovery has led to the use of these sequences as molecular markers to study gene linkage and mapping, genetic diversity, and evolution in addition to having applications in crop breeding and forensics (Hoshino *et al.,* 2012). SSR molecular markers also offer many benefits in their application for plant breeding and genetics due to their reproducibility, multi-allelic nature, codominant inheritance, relative abundance and good genome coverage (Powell *et al.,* 1996; Varshey *et al.,* 2005). Microsatellites are considered to be one of the most variable types of DNA sequences in genomes due to sequence length rather than primary sequence (Ellegren, 2004). A special quality of these markers is the higher mutation rate found when compared to the rest of the genome (Jarne and Lagoda, 1996; Oliveira *et al.* 2006). Microsatellites markers are highly transferable across certain taxa due to their highly conserved sequences (Peakall, 1998; Rico, 1996; Hoshino *et al*., 2012). In plants, SSR markers have been amplified since 1993 and are widely used for genetic diversity studies due to its ideal characteristics (Morgante and Olivieri, 1993; Varshey *et al.,* 2005; Wang *et al.,* 2009). For papaya, several microsatellite libraries have been produced and some of these markers have been shown to be transferable to other species in the *Caricaceae* family such as the *Vasconcellea* genera (Santos *et al*., 2003; Chen *et al*., 2007; Vidal *et al*., 2014; Ocampo *et al*., 2006; Sengupta *et al*., 2013). One of the most useful characteristics of SSR markers is their codominant nature which allows the differentiation between homozygote and heterozygote individuals.

Single Nucleotide Polymorphisms (SNPs) and Genotyping by Sequencing (GBS)

Single Nucleotide Polymorphisms (SNPs) are single nucleotide variations at a locus among individuals of the same species. SNPs were first discovered in human genomes and are considered to be the form on which the highest genetic variation can be identified among individuals of the same species, thus they are widely used as molecular markers (Mammadov *et al.*, 2012; Rafalski, 2002). Different from SSR markers, SNP variations within individuals is based on a single nucleotide difference at a particular

locus, therefore having a bi-allelic nature (Mammadov *et al*., 2012). Although SSR markers have been one of the preferred molecular markers to study plants, SNPs are becoming more popular due to their high abundance in the genome, automation for detection, and recently developed bioinformatics accessible analysis tools (Mammadov *et al.*, 2012). In plants, SNPs have been used for aiming to unravel the genetic diversity and population structure, genome wide association studies (GWAS), marker assisted selection (MAS), plant breeding, among others (Uitdewilligen *et al.,* 2013). SNPs can be identified by different methods such as using EST sequence data, array analysis, amplicon re-sequencing, and the most recent approach, genotyping by sequencing (GBS) (Ganal *et al.*, 2009; Rafalsky, 2002). GBS is a technique that implements Next Generation Sequencing (NGS) and has recently become a popular tool to identify Single Nucleotide Polymorphisms (SNPs) which then can be used to assess the genetic diversity and population structure of crops (Ganal, *et al.,* 2009). This technique consists on digesting genomic DNA with a restriction enzyme for further sequencing and then alignment of these sequences fragments to a reference genome leading to the identification of SNPs.

Papaya Genomics

A draft genome of papaya is available and was generated from a female plant of the "Sun Up" cultivar, the first PRSV transgenic resistant cultivar (Ming et al., 2008). The estimated coverage of this genome is 3X (Ming *et al*., 2008). The papaya genome size is 372Mbp and it is estimated to have a total of 24,746 genes. Sequences were generated using whole-genome shotgun sequencing (WGS) technique. Genome annotation is still partial to a super contig level and consisting of 3,207 scaffolds and 2,693 contigs in which 27,332 predicted genes are identified (Yu, *et al*., 2009, Vidal, *et al.*, 2014). Also, the papaya genome was evaluated for repetitive elements where Nagarajan *et al.* (2008), found that ~56% of the papaya genome is composed of repetitive sequences. Of these repetitive sequences transposons account for 52% while tandem repeats (microsatellites) account for 1.3% (Nagarajan *et al.*, 2008). Although a reference genome for papaya has been available for 9 years, the assessment of the genetic diversity of papaya using next generation sequencing technologies have not been reported to date.

CHAPTER 2 Literature Review

Genetic Diversity Studies of papaya

Genetic diversity is defined as any variation in nucleotides, genes, chromosomes, or whole genomes of organisms (Wang *et al.,* 2009). Several methods for studying genetic diversity of papaya have been used in the past. Early studies relied on morphological characteristics to evaluate a crop's diversity, which has limited resolving power due to the exclusion of characters that are not present in all developmental stages or that may be affected by environmental factors (Somasundaram and Kalaiselvam, 2011). In papaya, both morphological and molecular diversity studies have been performed in countries such as Costa Rica, Mauritus, Kenya, India, Venezuela, Brazil, Guadalupe and Barbados (Brown 2012; Madarbokus and Ranghoo-Sanmukhiya, 2012; Asudi *et al*., 2013; Sengupta *et al*., 2013; O'Campo, 2005; Matos *et al*., 2013). In these studies, different molecular markers have been used.

In 2002, Kim *et al.* (2002) evaluated the genetic diversity of papaya using AFLP markers. In this study, a total of 109 samples that included commercial varieties, hybrids, improved lines, and samples from different countries stored in the USDA collection were evaluated with 9 AFLP markers and the genetic similarity within samples was calculated. On average the genetic similarity was 0.880 with the highest value being 0.978 and the lowest 0.741. Interestingly, a value of 0.886 was calculated when analyzing the similarity among the USDA collection samples. This suggests that the samples from USDA assessed may represent the natural variation found in nature.

Another method used to assess papaya genetic diversity is the analysis of isozyme patterns. A study published by OCampo *et al.* (2006) revealed a low genetic diversity of papaya in the countries of Barbados, Guadeloupe, Grenada, Martinique, Trinidad, and Venezuela with the purpose of developing a papaya variety resistant to the bacterial pathogen *Erwinia papayae*. A total of 86 individuals and 9 isozyme patterns were assessed. With this data, the total diversity, heterozygosity (H_o) and F (Wright) Index was calculated (Table 4). A general tendency of low total diversity and low H_o was observed ranging from 0.27 to 0.42 and 0.18 to 0.42. Interestingly, low F Indexes were observed suggesting a discontinuous nature of the crop that leads to endogamy in very small populations and included a morphological assessment in which the genetic diversity analysis revealed a scattered dendogram with no geographical organization.

Table 4. Summary statistics for genetic diversity estimators. *Source*: OCampo (2006).

Population	Total Diversity	Heterozygozity	F (Wright)
Guadeloupe	0.38	0.28	0.23
Grenada	0.27	0.18	0.36
Martinique	0.30	0.38	-0.02
Trinidad	0.32	0.31	0.02
Venezuela	0.42	0.25	0.40
Barbados	0.41	0.42	0.01

A study by Sengupta *et al.* (2013) analyzed the genetic diversity of 41 samples from the *Caricaceae* family using 20 SSR markers. The samples' provenance was mainly from India and included accessions from the genus *Vasconcellea, Jacaratia*, and *Carica*. This research group focused their analysis in calculating the polymorphic information content (PIC) for the markers used which on average was 0.73. All the markers used were polymorphic and an average of 7 alleles per locus was reported. A dendogram based on Jaccard's similarity coefficient with a bootstrap of 1000 permutations, revealed that the samples appeared to be very distinct, showing inherent genetic diversity with the highest and lowest similarity among samples being 67% and 7%, respectively. The main explanation for this finding was the existence of a high gene flow among the Indian accessions along with the long history of the crop in India (Prest, 1955). Brown *et al.* (2012) compiled data from 184 papaya samples: 164 of them belonging to Costa Rica's natural papaya populations (10 operational populations) and 20 different accessions from the USDA papaya germplasm collection. Contrary to Sengupta *et al.* (2013) a low genetic diversity was found, reporting H_o values that ranged from 0.14 (USDA accessions) to 0.45 in the natural population. The observed allele per locus varied between 6 and 25 with a mean of 11.6 alleles per locus. Brown *et al.* (2012) also assessed the population structure and reported that the most probable k (number of groups) for the assessed samples were k=2 and k=3. A study by Asudi *et al.* (2013), also evaluated papaya genetic diversity in Kenyan germplasm. For this assessment 42 accessions from Kenya were evaluated with 7 SSR markers. A high genetic similarity was reported with values ranging from 0.764 to 0.932 with a mean of 0.844. The authors reported that the samples used in his study were closely related although a mean H_o of 0.62 was calculated. Also, Matos *et al.* (2013) used 15 SSR markers and evaluated 96 papaya accessions

from Brazil. Matos *et al.* (2013) reported a total of 68 alleles with a range of 3-10 alleles per locus. The mean H_o was 0.20 and F-index of 0.58. The population structure had predicted k value equal to 6.

CHAPTER 3 Genetic Diversity of Papaya (*Carica papaya*) using SSR Markers

Summary

The genetic diversity of papaya was evaluated using 23 SSR markers. A total of 139 samples from Puerto Rico were compared to 13 varieties from the USDA germplasm repository and 10 commercial varieties. A high allelic abundance but low observed heterozygosity (H_o) was recorded.

Methodology

Plant Material

A total of 162 samples were evaluated (Appendix Table 1). Of these samples 139 were unknown accessions from different municipalities of Puerto Rico, which were acquired voluntarily from Puerto Rico habitants' personal gardens. These unknown samples were grouped into 5 groups based on their geographical location (Figure 2). Further 13 samples from the USDA germplasm repository in Hawaii and 10 samples were acquired commercially. Leaf material was collected and frozen until further analysis. The USDA germplasm accessions and commercial varieties were planted from seeds and leaves were collected approximately 15 weeks after planting.

Figure 2. Sample distribution based on location. The 139 unknown samples were grouped based on their location in Puerto Rico. North West (red), North East (purple), South East (green), South West (blue), Central (yellow).

DNA Extraction

DNA was extracted from papaya leaves using a modified protocol based on Doyle and Doyle (1991). Approximately 0.5g of leaf was ground with sterile sand using a pestle in a 2.0 mL tube. After that, 800 μL of 3% CTAB buffer [20 mM EDTA, 0.1 M Tris-HCl pH 8.0, 1.4 M NaCl, 3% CTAB, 3% PVP, 0.2% β-mercaptoethanol] was added and shaken by inversion. Samples were incubated at 70°C for 30 minutes. Subsequently 500 μL chloroform: isoamyl alcohol [25:1] was added and gently mixed by inversion. The samples were centrifuged for 3 minutes at 13,200 rpm and a total of 500 μL of the supernatant were transferred to a new 2.0 mL tube. An equal amount of chloroform: isoamyl alcohol [25:1] and 200 μL of 3% CTAB buffer was then added. After mixing by inversion, samples were centrifuged for 3 minutes at 13,200 rpm. The supernatant was transferred to a 1.5 mL tube and 350 μL of cold isopropanol (-20°C) was added. After gently mixing by inversion, samples were centrifuged for 5 minutes at 13,200 rpm. Afterwards, the supernatant was discarded and the pellet was left to dry for approximately 5 minutes. Then, 700 μL of cold (-20°C) 70% ethanol was added and the tubes were centrifuged for 3 minutes at 13,200 rpm. The clean pellet was dried at room temperature for 5 minutes. DNA pellet was resuspended in 200μL of $T_{10}E_1$, and 4 μL of RNAse was added prior to incubation at 65°C for 5 minutes. DNA samples were quantified using a NanoDrop Lite Spectrophotometer (Thermo Scientific Inc., Wilmington, DE, USA), diluted to a final concentration of 25ng/μL and stored at -20°C until further use.

PCR Reaction

Twenty-three SSR markers were selected (Table 5) for this study taking in consideration a high Polymorphic Index Content (PIC) reported in previous studies. A master mix was prepared for individual PCR reactions of 25 μL each. Final concentration of the reagents per reaction were: 1X Promega Colorless Gotaq Flexi Buffer, 2mM Promega $MgCl_2$, 2mM of KCl, 2mM Tris-HCl pH 8.2, 0.4mM Promega dNTPs, 0.4μM of forward and reverse primers each (Table 5), 0.5pm/μL of LI-COR fluorescently labeled M13 IRDye forward primer (LI-COR Bioscience, Lincoln, NE, USA), 1 unit of BioReady rTaq DNA Polymerase (Bulldog Bio, Portsmouth, NH, USA), and 40ng of template DNA. The PCR cycle consisted of an initial denaturing step at 94°C for 5 minutes, 30 repetitions of the following cycle 94°C for 1 minute, 30 seconds at 50°C for annealing, and a 72° extension of 1 minute. A final extension step of 72°C for 5 minutes was added followed by infinite hold at 4°C. In this method of PCR, a fluorescently tagged PCR product is obtained due to the annealing of the fluorescently labelled M13 IRDye primer to each of the SSR forward primers. Each of the SSR forward primers were designed with a M13 sequence overhang at the 5' end

which facilitates the annealing of the M13 primer. The PCR Product was diluted 1:5 in LI-COR blue stop solution, denatured at 94°C for 5 minutes and electrophoresed in a 6.5% Polyacrylamide Gel. The gel was prepared using 20 mL of LI-COR 6.5% polyacrylamide gel matrix, 150 μL of 0.01% Ammonium Persulfate, and 15 μL of TEMED. Electrophoresis was performed with default settings on LI-COR 4300 (LI-COR Bioscience, Lincoln, NE, USA).

Data Analysis

Alleles were scored by visual inspection with respect to molecular markers and stutter bands were accounted for. A weight matrix was constructed using Microsoft Excel software. Genetic Diversity estimators (H_o, H_E, F Index, and number of private alleles) were calculated using GenALEx 6.5 (Peakall and Smouse, 2012). Polymorphic Index Content (PIC) and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) dendogram was generated by Eucledian genetic distance using PowerMarker Software V. 3.25 (Liu *et al*., 2005). Dendogram visualization was achieved using Interactive Tree of Life (iTOL) software (Letunic and Burk, 2007). The genetic structure of the samples was evaluated using STRUCTURE software using a burning length of 20,000 and subsequent 100,000 repetitions after burn-in. The number of clusters (k) was evaluated from 1 to 10 with 5 iterations and the most probable k was identified using Structure Harvester Web v0.6.94 (Earl and vonHoldt, 2012) by the Evanno (2005) method of delta k (Δk). For the Puerto Rico unknown samples, the number of clusters (k) was evaluated from 1 to 20 with 6 iterations.

Table 5. List of 23 used SSR markers, primer sequences, motif, allele sizes, observed heterozygosity (Ho) and expected heterozygosity (Ht) per locus.

Locus	Primer Forward	Primer Reverse	Motif	Allele Size	H_0	Ht
AJ810489	GTCTATCTAC CTCCCA	GAGTGTTATC ATAGTCTACA	(TC)24	259-295	0.25	0.81
AJ810490	GAACTCACCT ACACGAACT	ACTTCTACCAC CGGC	(TC)14	202-236	0.33	0.68
AJ810491	AAGCCAAGAA CAGCAA	AATGCTTGAA GTAAACACC	(TC)10	239-253	0.22	0.72
AJ810492	GCATTACTTA TCATCGTCC	ACTATCCTTG GCGTCTT	(CT)18	588-603	0.22	0.67
AJ810493	CCAAAACGGA AAACAC	ATCAAGCTCC CTTTCAC	(TG)10(AG)7(GA)10	288-303	0.006	0.70
AJ810494	CCAACACATT CATCCAC	CTGAAGCATT ACCGAGA	(TC)18	239-253	0.48	0.78
AJ810495	ATGGCTGAAG ACAACTC	CTCAATAGCC CAATAACA	(CT)20…(AC)5	306-322	0.11	0.51
AJ810505	ATGGGATTTT AGAGGTG	GTATGAGGGA ATGGAAA	(CT)9…(CT)9	312-318	0.17	0.50

Cont. Table 5. List of 23 used SSR markers, primer sequences, motif, allele sizes, observed heterozygosity (Ho) and expected heterozygosity (Ht) per locus.

Locus	Primer Forward	Primer Reverse	Motif	Allele Size	H_0	Ht
CP21	ATCGACCGAG GAAGGTACG	TCAAAAACCC ATTGAGTCTG C	(GT)21	152-207	0.15	0.59
CP31	AAGGGTACGT CATGGAGCA	TCTGTCGCCTT TTATACTCTTG	(AT)6(GT)10	167-178	0.26	0.87
CP44	TGACAACGAA CTACATCCCT A	CCTCATGGTTT GTGTACTCCT	(AT)12	244-267	0.33	0.67
CP49	CCTGAAAGCA ACCATTTCTA	TCGCTGGAGC TGTAAGAGA	(AT)12	210-222	0.36	0.66
CPCIR2	GGTCTTTTAG TTCCAGTGTT	ATGATTGAGC GGGTG	(GA)12	252-294	0.26	0.86
CPCIR3	CGCATTGTTA TTGACT	ACCTACAGGG CCTAC	(TC)8	203-234	0.17	0.70
SP1	TGCAACAGAA ATAAAAACAG CA	GACGTGGACG AGCTCTGTGT	(TTTC)5/(TTC)9	273-473	0.006	0.71
SP3	CACCAACAAG TTCCTTGGGT	TGCATGCATG TGTGTGGATA	(AC)9	648-700	0.08	0.62
SP4	TGCTCATAAA GTGATGGAG GT	TGGCGACCAT TTAAACAACA	(AT)9	90-200	0.06	0.74
SP5	TTGGCTTCAA ATTCAGGCTT	GCGGCTTCTG GATCTGATAA	(AC)9	242-267	0.03	0.68
SP6	CTTGCACCGA ACCCTAAAAG	CATGAAAAAC ACATGCCTGC	(AT)9	675-690	0.22	0.54
SP7	CAGTTGTAGG GGTTGGTGGT	GTCCACAAAT CAGAGCCCAT	(AAT)7	310-315	0.22	0.47
SP8	CAAATCATGT TGGTCTGCGT	GCTCAGCGGC TATTTTTGAC	(ATT)7	355-422	0.50	0.65
SSPA3	CGAAGCAAAA CTTCTCAGCC	TCTCAATTTCC ATTTTCCGC	(AG)10	155-177	0.11	0.66
SSPA8	TGTCTCAGCA TATCCACCCA	ATGGCCTTTT GGAACATCAG	(AT)12	588-604	0.08	0.30
Mean					0.20	0.66

Results

Genetic Diversity

A total of 214 alleles were observed across the 23 SSR markers. Allele per locus ranged between 2 to 18 with a mean of 9.3 (Figure 3). All evaluated loci were polymorphic with its Polymorphic Index Content (PIC) ranging from 0.292 (locus AJ810493) to 0.863 (CPCIR2) with a mean of 0.626 (Table 7). A

low observed heterozygosity (H_o) for all individually evaluated loci was obtained with a mean of 0.2047. Values (H_o) values per locus ranged from 0.00617 (AJ810493 and SP1) to 0.506 (SP8). The mean observed heterozygosity for all the samples is 0.219 (Table 8) and the mean inbreeding coefficient (F-index) is 0.565. When evaluating Puerto Rico Unknown Samples by different geographical regions, the observed heterozygosity (H_o) ranged from 0.196 in the Northwest Area to 0.284 in the North East area. The observed heterozygosity for the USDA germplasm samples was 0.138 (Table 8). The USDA germplasm group and known commercial varieties group showed to have the greatest amount of private alleles with 22 and 20, respectively, whereas a total of 26 private alleles where found among the Puerto Rico unknown samples (Table 6).

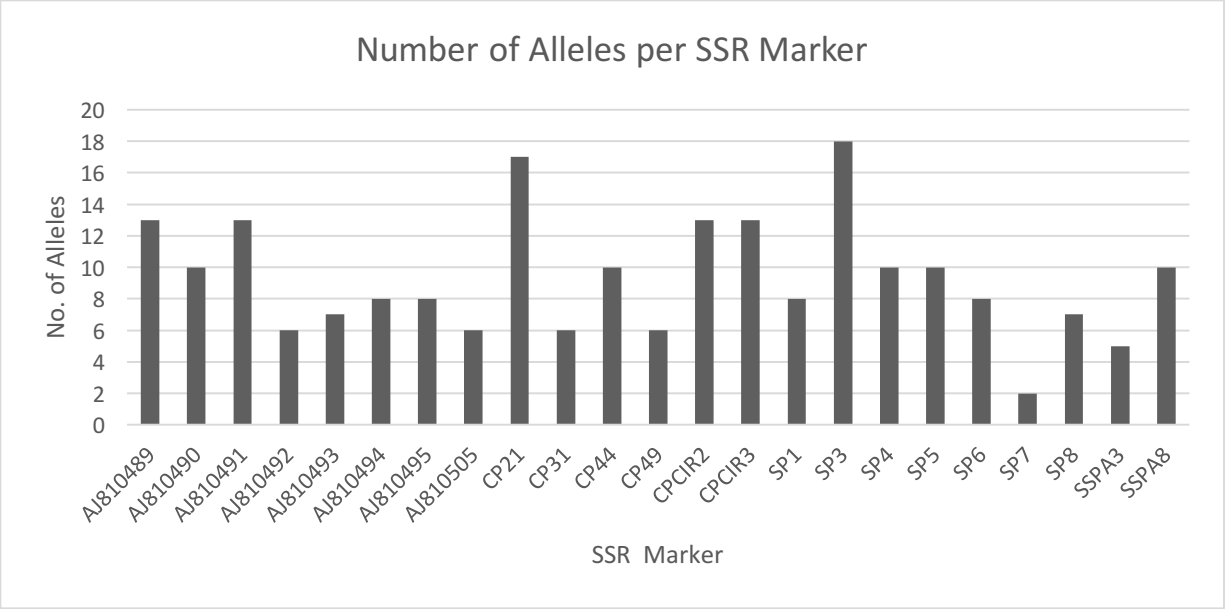


Figure 3. Number of alleles recorded per SSR marker for 162 papaya samples consisting of 139 unknown accessions from Puerto Rico, 13 accessions from USDA germplasm collection and 10 known commercial varieties. Allele number ranged from 2 to 18 alleles per locus with a mean of 9.304.

Table 6. Number of Private Alleles per assessed groups.

Population	Number of Private Alleles	Percentage (%) of Private Alleles per Population	Percentage (%) of Private Alleles (Total)
Puerto Rico Unknown NW	3	3.13	1.40
Puerto Rico Unknown NE	7	7.29	3.27

Cont. Table 6. Number of Private Alleles per assessed groups

Population	Number of Private Alleles	Percentage (%) of Private Alleles per Population	Percentage (%) of Private Alleles (Total)
Puerto Rico Unknown SE	4	4.00	1.87
Puerto Rico Unknown Center	8	9.30	3.74
Puerto Rico Unknown SW	4	4.55	1.87
USDA	22	21.36	10.28
Commercial	20	26.32	9.35

Table 7. List of analyzed locus with the Number of Alleles recorded (Na), Number of Effective Alleles (Ne) and Polymorphic Index Content (PIC) for each. Values calculated by analysis of 162 papaya samples with 23 SSR markers using PowerMarker software.

Locus	**Na**	**Ne**	**PIC**
AJ810489[a]	13.000	4.568	0.79
AJ810490[a]	10.000	2.740	0.64
AJ810491[a]	13.000	2.713	0.60
AJ810492[a]	6.000	2.084	0.60
AJ810493[a]	7.000	1.239	0.29
AJ810494[a]	8.000	2.928	0.68
AJ810495[a]	8.000	2.601	0.64
AJ810505[a]	6.000	2.596	0.65
CP21[b]	17.000	4.210	0.77
CP31[b]	6.000	1.846	0.48
CP44[b]	10.000	1.819	0.49
CP49[b]	6.000	2.080	0.55
CPCIR2[a]	13.000	7.025	0.86
CPCIR3[a]	13.000	2.778	0.64
SP1[c]	8.000	2.474	0.62
SP3[c]	18.000	6.221	0.84

Cont. Table 7. List of analyzed locus with the Number of Alleles recorded (Na), Number of Effective Alleles (Ne) and Polymorphic Index Content (PIC) for each. Values calculated by analysis of 162 papaya samples with 23 SSR markers using PowerMarker software.

Locus	**Na**	**Ne**	**PIC**
SP4[c]	10.000	2.985	0.70
SP5[c]	10.000	3.037	0.67
SP6[c]	8.000	2.622	0.63
SP7[c]	2.000	1.882	0.46
SP8[c]	7.000	1.637	0.44
SSPA3[c]	5.000	3.109	0.65
SSPA8[c]	10.000	2.029	0.59
Mean	9.304	2.923	0.62

[a] OCampo *et al.* (2006)
[b] de Oliveira *et al*. (2010)
[c] Sengupta *et al*. (2013)

Table 8. Genetic Diversity Estimators and standard error for the 162 assessed samples and groups. Number of alleles (Na), Observed Heterozygosity (Ho), Expected Heterozygosity (He), and Inbreeding Coefficient (F).

Samples		**Na**	**Ho**	**He**	**F**
Puerto Rico Unknown	Mean	7.174	0.231	0.559	0.576
	SE	0.628	0.031	0.031	0.050
Puerto Rico Unknown (Groups)					
Puerto Rico Unknown (NW)		4.174	0.196	0.494	0.579
	SE	0.331	0.031	0.031	0.063
Puerto Rico Unknown (NE)		4.043	0.284	0.532	0.431
	SE	0.336	0.034	0.033	0.074
Puerto Rico Unknown (SE)		4.348	0.257	0.516	0.502
	SE	0.353	0.038	0.031	0.064
Puerto Rico Unknown (Center)		3.739	0.238	0.492	0.527
	SE	0.334	0.033	0.029	0.065
Puerto Rico Unknown (SW)		3.826	0.209	0.536	0.614
	SE	0.272	0.038	0.036	0.060

Cont. Table 8. Genetic Diversity Estimators and standard error for the 162 assessed samples and groups. Number of alleles (Na), Observed Heterozygosity (Ho), Expected Heterozygosity (He), and Inbreeding Coefficient (F).

Samples		**Na**	**Ho**	**He**	**F**
USDA germplasm	Mean	4.087	0.138	0.582	0.787
	SE	0.371	0.045	0.033	0.064
Commercial Varieties	Mean	4.043	0.288	0.549	0.510
	SE	0.347	0.063	0.042	0.105
Mean over Loci and Groups					
		Na	**Ho**	**He**	**F**
Total	Mean	5.101	0.219	0.563	0.565
	SE	0.319	0.028	0.021	0.046

UPGMA Dendogram

A genetic distance method based dendogram was constructed using Euclidean distance and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Figure 4). Two main clusters were identified and no geographical grouping among the samples was identified. Samples from the USDA germplasm and commercial accessions, grouped within the same cluster but not exclusively since samples 'Mona', 130 (SE) and 67 (SW) grouped within the same cluster as the commercial and USDA varieties. Cluster 1 is composed 7 samples from the Puerto Rico Unknown group from the North East (2), North West (2), Southeast (1) and Southwest areas (2) while cluster 2 included majority of the unknown Puerto Rico samples, all of the commercial varieties and USDA germplasm samples.

STRUCTURE Analysis

When analyzing the population structure of the 162 samples, a total of 2 clusters (k=2) were identified after using the method of delta k with Structure Harvester software (Figure 5). All the Puerto Rico unknown samples but one, the 'Mona' sample, were found to belong to cluster 1 (Figure 6). Cluster 2 comprised of the USDA germplasm samples and the known commercial varieties samples. Limited admixture was observed among the two identified clusters with few samples showing the following parameters: Sample 37s showed 0.850 inferred ancestry level of cluster 2 and 0.150 of cluster 1, and

sample 79s showed 0.409 inferred ancestry level of cluster 2 and 0.592 ancestry level of cluster 1. For the population structure analysis of the Puerto Rico Unknown samples, a total of 2 clusters (k=2) were identified after using the delta k method by Evanno (2005) (Figure 7). No geographical correlation was identified upon the samples estimated ancestry and the positioned cluster (Figure 8).

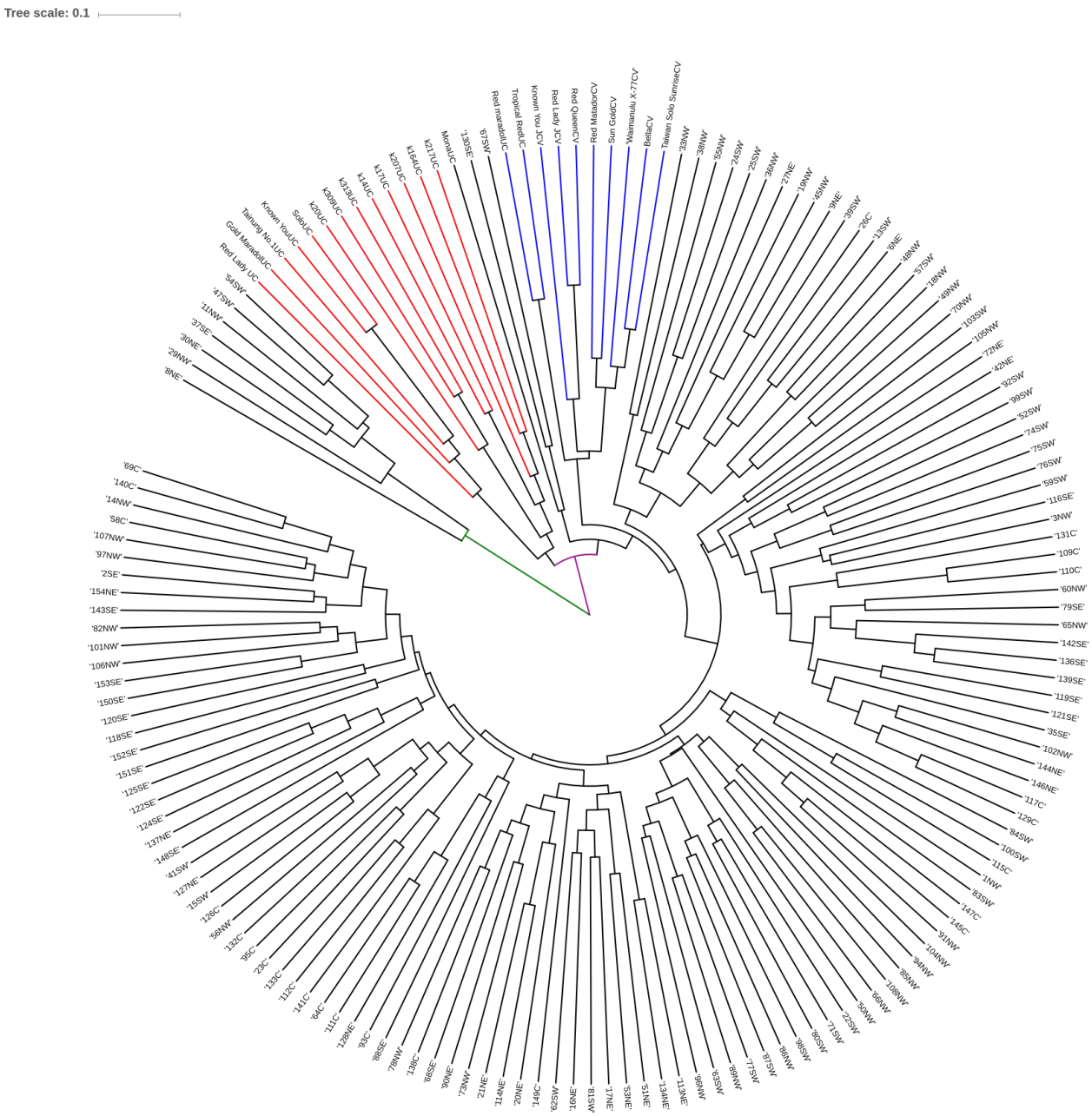


Figure 4. 162 sample UPGMA dendogram using Euclidean distance.

Cluster 1 (green node), cluster 2 (purple node). 139 Unknown Puerto Rico Samples (black), 13 USDA germplasm collection samples (red), 10 commercial varieties (blue).

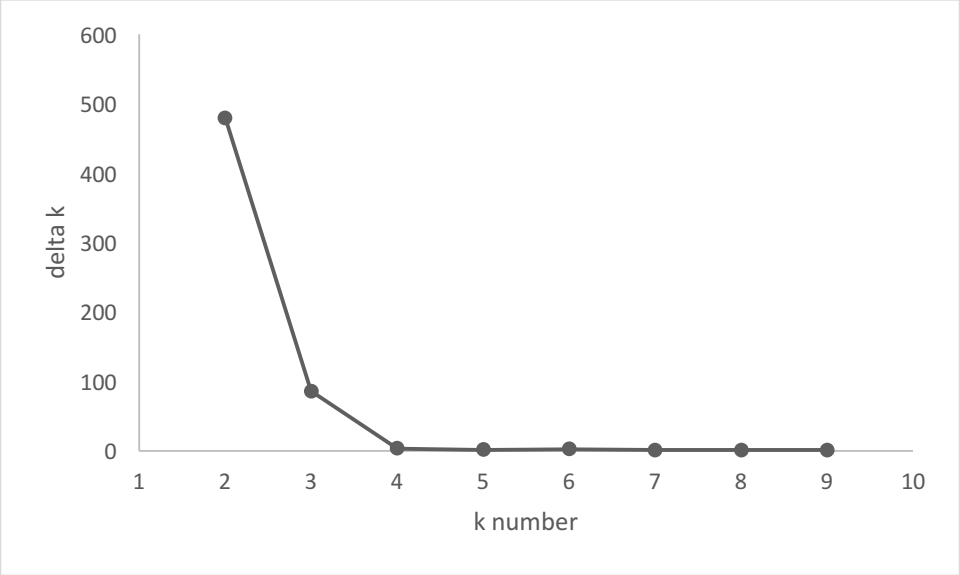


Figure 5. Delta k value by number of k (groups) calculated using Structure Harvester software for the 162 evaluated samples.

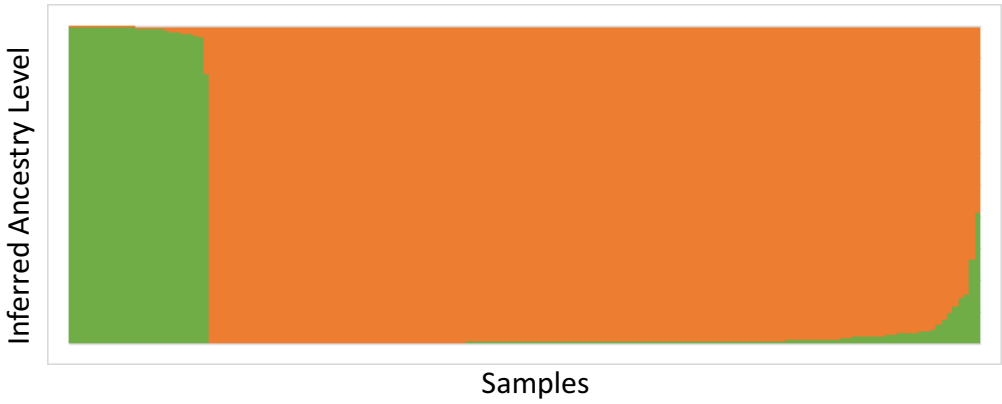


Figure 6. Population structure bar plot.

Population structure bar plot shows two groups (k=2) identified by delta k method for the 163 sample analysis. Group 1 (orange) is composed of the unknown samples from Puerto Rico and group 2 (green) is composed of USDA germplasm samples, known commercial varieties, and a sample from the Puerto Rican island 'Mona'.

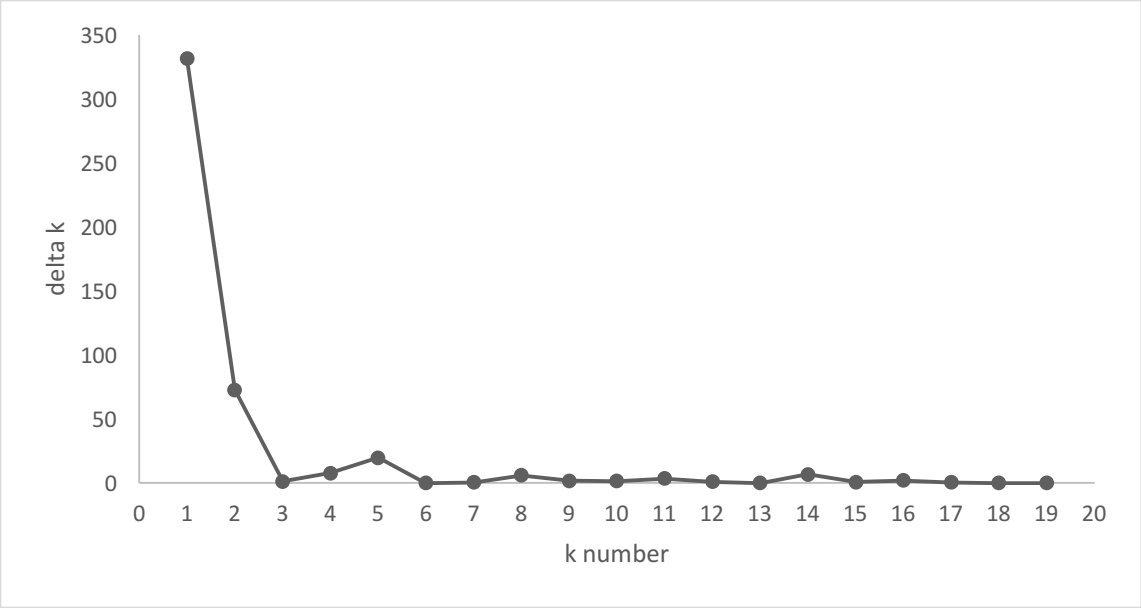


Figure 7. Delta k value by number of k (groups) calculated using Structure Harvester software for the Puerto Rico Unknown samples.

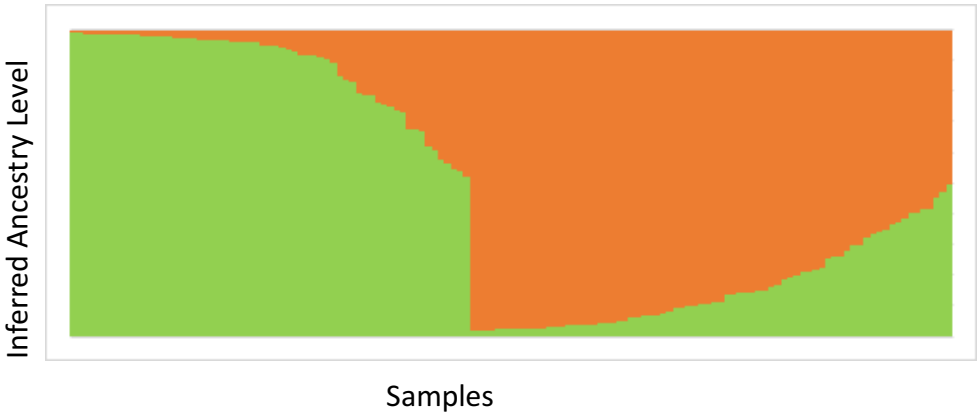


Figure 8. Puerto Rico Unknown Samples population structure barplot.

Population structure bar plot shows two groups (k=2) identified by delta k method for the Puerto Rico Unknown sample analysis. Group 1 is shown in orange and group 2 in green.

Discussion

Genetic diversity estimators revealed similar results to other genetic diversity studies using SSR markers (Matos *et al.,* 2013; Brown *et al.,* 2012; Asudi *et al.,* 2013). When evaluating H_o for the USDA germplasm samples, we found similar results as Brown *et al.* (2012) and Rieger (2009) which evaluated a total of 20 samples from the USDA germplasm repository and reported an observed heterozygosity H_o of

0.14 (similar to 0.138 in our analysis). We also found low levels of heterozygosity in Puerto Rico unknown samples with an observed heterozygosity ranging from 0.196 to 0.284 which was comparable with Costa Rican natural populations that ranged from 0.31 to 0.45 (Brown *et al.*, 2012). Likewise, Matos *et al.* (2013) also reported low levels of heterozygosity when evaluating a Brazilian germplasm that resulted in a mean of 0.20. Regarding the total allele number, our study showed a high number of alleles (214) when compared to other studies (OCampo *et al.* 2007; Matos *et al.*, 2013; Asudi *et al.*, 2013) but similar to Rieger (2009) and Brown *et al.* (2012). Nevertheless, allele per locus in our study which ranged from 2-18 with a mean of 9.304 is comparable with other studies such as Asudi *et al*. (2013) that reported 8 to 18 alleles per locus with a mean of 11.93, and Brown *et al.* (2012) that reported 6 to 25 allele per locus with a mean of 11.6.

A possible explanation for high allelic abundance may be the history of papaya in Puerto Rico. Although the history of papaya in Puerto Rico is not well documented, we infer that multiple introductions have been made due to the localization and political history of the island. It is thought that papaya was first introduced to Puerto Rico around 1525 due to the proximity and likewise history to the Dominican Republic where the introduction of papaya has been documented to have occurred around 1525 (Teixeira da Silva *et al.,* 2007). Since Puerto Rico does not produce enough papaya to meet the local demand, papaya fruits are regularly imported from the Dominican Republic, Costa Rica and the United States of America (Morton 1987; Zambrana-Echevarría *et al.,* 2016; Junta de Planificación Puerto Rico, 2016). This undoubtedly contributes to more allelic diversity across the island as papaya grows from seeds and is easily cultivated for personal consumption by residents of Puerto Rico. For example, during 2015, Puerto Rico imported 512,861 kg of papaya from Costa Rica and interestingly exported 50,072 kg to the United States of America (Junta de Planificación Puerto Rico, 2016). Another possible introduction event of papaya to Puerto Rico was during 1978, when a new economic development strategy was implemented by establishing agriculture as one of the pillars for an export-based economy. This led to different approaches with one of them being converting the southern coastal area of Puerto Rico as an intensive fruit and vegetable farming area for local consumption and winter exportation (Carro-Figueroa, 2002).

Structure analysis revealed a total of 2 distinct groups. One of the groups contains 138 unknown Puerto Rico samples with the exception of a sample from Mona island, an uninhabited island belonging to the archipelago of Puerto Rico. Several rationalizations for the distinctive characteristic of the "Mona" sample may be described. Wadsworth (1972) reported crop cultivation of cassava and melons in Mona island since the Taino period on the island which is officially dated since Mona's discovery during 1493 or 1494. Wadsworth also describes the farming of 32 acres of different crops including papaya during 1922

(Wadsworth, 1972). More evidence on the history of papaya in Mona is found in a botanical survey of Mona performed on 1914 were *Carica papaya* was reported to be present in the coastal plain and to be presumably established after cultivation (Britton, 1915). Another possible explanation for this sample being more genetically similar to the known commercial and USDA germplasm is that it is actually a known cultivar that has remained isolated and self-fertilized therefore more similar to known samples. Mass human migration from the Dominican Republic to Puerto Rico is documented since 1961, when political events such as the fall of Trujillo's regime and consequent events lead to Puerto Rico being a preferred destination due to its proximity, similar history, geography, culture and language (Duany, 2005). This could possibly explain another introduction of papaya to this uninhabited island since it is known that migrants attempt to cross the Mona passage in order to access Puerto Rico main island (US Coast Guard, 2016).

The fact that the UPGMA dendogram does not match identically to the Structure analysis may be due to the difference in clustering methods; UPGMA method is distance based and Structure method utilizes Bayesian inference (Evanno *et al.*, 2005). Nevertheless, these analyses were similar. The UPGMA dendogram have several clusters one of them containing the known commercial samples and the USDA germplasm samples but also containing samples from the SE (130) and SW (67). We believe human transportation of papaya seeds as the reason for the lack of geographical clustering within the unknown samples of Puerto Rico. It is known that plant distribution and diversity is influenced by human behavior due to the mobility capacity of such (Niggerman *et al.*, 2009; Antrop, 2004). We had duplicates of the samples 'Known You' and 'Red Lady' acquired and classified by different means. One sample each of these varieties was acquired with the USDA repository samples while the other was acquired commercially. Interestingly, these samples were not identified as clones.

In general, we suggest that Puerto Rico is an allele reservoir for papaya that could be further studied for possible breeding applications given the allelic abundance the island has. We believe that the abundance is due to historical reasons specifically due to the geographic location of Puerto Rico which is central and accessible to all American continent (Zambrana-Echevarria *et al.*, 2016). We suggest future studies evaluating the genetic diversity at a morphological level taking in consideration the allele abundance in the assessed samples. This study provides the first exhaustive record of the genetic diversity in Puerto Rico that can be utilized in conservation or breeding purposes.

Conclusions

Conclusions

- This is the first exhaustive record of the genetic diversity of *Carica papaya* in Puerto Rico.

- Low observed heterozygosity (H_o) values and high F-index were recorded in this study and can be explained by the human population preference to the hermaphrodite form (and consequently self-fertilization) of papaya.
- Two groups were identified when assessing the population structure of all the assessed samples and the Puerto Rico Unknown samples.
- The distribution of the samples in the UPGMA dendogram can be explained by the historical, social, and commercial movement of papaya seeds in Puerto Rico.

Recommendations

- To include more commercial and USDA germplasm varieties in the assessment, thus having more representation of the existing global genetic variation.
- To include more SSR markers in the study.
- To include more samples from natural populations of papaya in Puerto Rico.

CHAPTER 4 Genetic Diversity of papaya (*Carica papaya*) using Genotyping by Sequencing (GBS)

Summary

The genetic diversity of papaya was evaluated using SNPs identified by GBS. A total of 131 samples from Puerto Rico were compared to 13 varieties from the USDA germplasm repository and 10 commercial varieties. Low observed heterozygosity (H_o) and low Gis index were recorded.

Methodology

Plant Material

Samples used for this experiment were the same used in the SSR analysis (see chapter 3, Appendix Table 1).

DNA Extraction

DNA was extracted according to a protocol developed by Hugo Cuevas Laboratory of the USDA-Tropical Agriculture Research Station, Mayaguez PR. (personal communication) was used. Approximately 0.5g of leaf tissue was ground using 2 mL of extraction buffer [100mM Tris HCl pH 8.0, 50mM EDTA pH 8.0, 500mM NaCl, 2% SDS, 1% PVP-360] in a mortar and pestle. Homogenate was transferred to a 2 mL sterile tube. Then, samples were incubated with 4 μL of RNAse and mixed by inversion every 10 minutes at 65°C for 30 minutes. After incubation, 150 μL of 5M Potassium Acetate pH 6.5 were added and samples were incubated 5 minutes at 4°C. Subsequently, 500μL of chloroform: isoamyl alcohol [24:1] were added before centrifugation at 13,200 rpm. The supernatant was transferred to a sterile 1.5 mL tube and 400 μL of cold isopropanol were added. Samples were incubated for 15 minutes at -20°C and centrifuged at 13,200 rpm. The supernatant was discarded and the pellet was dried at room temperature for 5 minutes before suspending it in 50 μL of autoclaved ddH_2O.

DNA Purification

DNA samples were purified using ZR-96 DNA Clean & Concentrator™-5 kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions.

Genotyping by Sequencing (GBS)

Purified DNA samples were sent to the Genomic Diversity Facility at Cornell Institute of Biotechnology (Cornell University, Ithaca, NY, USA) for Genotyping by Sequencing, a method developed by Elshire *et al*., 2015. For sample preparation, genomes were digested with restriction enzyme ApeKI.

Data Analysis

Genotyping by Sequencing raw data was analyzed using different software. TASSEL v5.0 software (Bradbury *et al.*, 2007) was used to filter samples with missing data, filter SNPs by frequency, constructing an Identity by State (IBS) distance based Unweighted Pair Group Method with Arithmetic Mean (UPGMA) dendogram, and obtaining SNP position list. The UPGMA dendogram was visualized using Interactive Tree of Life (iTOL) software (Letunic and Burk, 2007). Linkage Disequilibrium (LD) pruning was performed with Plink v1.9 software (Chang *et al.*, 2015) using a sliding window of 50 SNPs, 5 steps, and R^2<0.5 as parameters. The genetic diversity estimators were obtained using Genodive software version 2.0b27 (Meirmans and Van Tienderen, 2004). After data filtering, STRUCTURE software was used to evaluate the population structure of the samples. A burning length of 20,000 was used and subsequent 100,000 repetitions after burn-in. The number of clusters (k) was evaluated from 1 to 10 with 5 iterations and the most probable k was identified using Structure Harvester Web v0.6.94 (Earl and vonHoldt, 2012) by the Evanno (2005) method of delta k (Δk). When evaluating the Puerto Rico unknown samples, the number of clusters (k) was evaluated from 2 to 20 with 3 iterations.

Results

Genetic Diversity

A total of 16,170 SNPs were genotyped and aligned to the papaya reference genome obtained from the United States of America Department of Energy Joint Genome Institute Phytozome database v11. After data filtering to eliminate non-informative and redundant data, a total of 4,245 SNPs were selected. Complete subsequent analysis was performed for SNPs in a frequency of 0.05-0.95 (Table 9) and eliminating those in Linkage Disequilibrium (LD) with an R^2 greater than 0.5 (LD pruning) (Appendix Table 2). Also, samples that failed to be genotyped, were eliminated from the analysis. The mean observed heterozygosity (H_o) was 0.226 and the inbreeding coefficient (Gis) 0.067 (Table 10). Within the assessed groups of samples, H_o ranged from 0.160 (USDA germplasm samples) to 0.275 (PR-NE). The inbreeding

coefficient for all of the groups except the USDA germplasm samples were low ranging from -0.105 (PR-NW) to -0.605 (commercial varieties). The USDA germplasm samples exhibited a Gis of 0.128. Filtered SNPs were identified to be located in 415 contigs/supercontigs of the partially annotated reference genome.

Table 9. Single Nucleotide Polymorphism (SNPs) Frequency before LD pruning.

SNP frequency	**Amount of SNPs**
Raw	16, 170
0.01-0.99	16, 155
0.05-0.95	7,990
0.10- 0.90	6,314

Table 10. Genetic Diversity Estimators for the 154 assessed samples.

Observed Heterozygosity (H_o), Expected Heterozygosity (He), and Inbreeding Coefficient (Gis).

Samples		**Ho**	**He**	**Gis**
Puerto Rico Unknown	Mean	0.236	0.216	-0.092
Puerto Rico Unknown (Groups)				
Puerto Rico Unknown (NW)		0.233	0.211	-0.105
Puerto Rico Unknown (NE)		0.275	0.199	-0.382
Puerto Rico Unknown (SE)		0.257	0.209	-0.229
Puerto Rico Unknown (Center)		0.270	0.192	-0.405
Puerto Rico Unknown (SW)		0.233	0.191	-0.218
USDA germplasm	Mean	0.160	0.183	0.128
Commercial Varieties	Mean	0.252	0.157	-0.605
Mean over Loci and Groups		**Ho**	**He**	**Gis**
Total	Mean	0.226	0.242	0.067
	SD	0.005	0.003	0.027

UPGMA

An UPGMA dendogram was generated using an IBS distance matrix constructed with TASSEL v5.0 software (Bradbury *et al.*, 2007). Two main clusters were identified, one which have an out group appearance due to its only sample present "Mona" (Figure 9). Cluster 2 contains all the Puerto Rico unknown samples, the commercial varieties, and the USDA germplasm collection varieties. Commercial varieties and USDA germplam samples were distributed among different but close sub clusters. Interestingly, samples "k14", "Tainung No. 1", and "Bella", positioned among the majority of the Puerto Rico unknown samples. No geographical clustering within the Puerto Rico unknown samples was apparent.

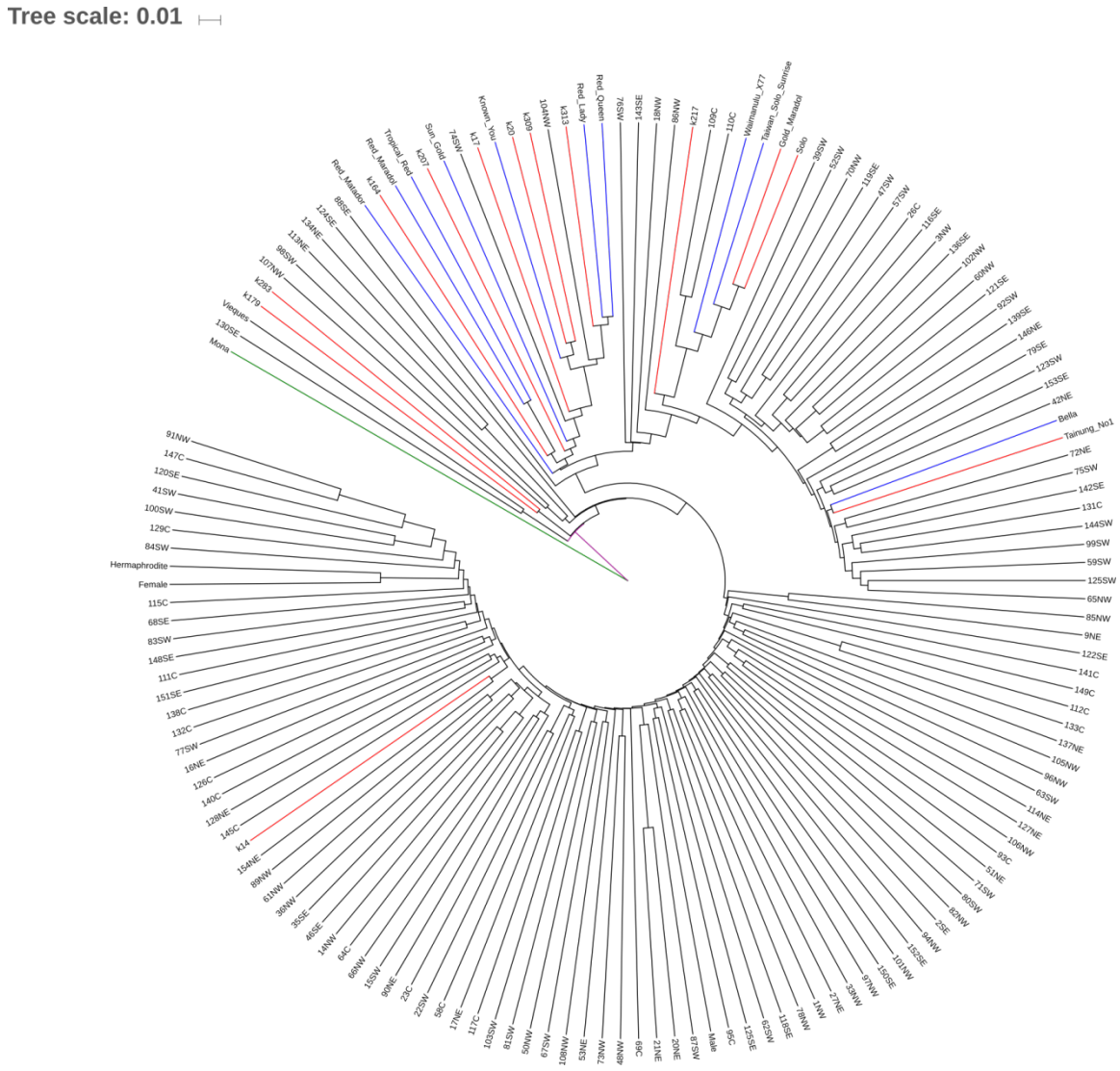


Figure 9. 154 sample identity by state distance based UPGMA dendogram.

Cluster 1 (green node), cluster 2 (purple node). 132 Unknown Puerto Rico Samples (black), 12 USDA germplasm collection samples (red), 10 commercial varieties (blue).

STRUCTURE

A total of 2 groups (k) were identified using the most probable k method of Evanno (2005) (Figure 10). The barplot (Figure 11) shows the 154 assessed samples in 2 identified clusters in green and red. The clusters are composed of different samples with no geographical grouping observed. Interestingly, the samples "Red Queen" and "Sun Gold" show a complete inferred ancestry (1.0) from cluster 1 and 2 respectively. When evaluating the population structure for only the unknown Puerto Rico samples, a total of 7 groups (k) were identified by the most probable k method of Evanno (2005) (Figure 12). This data suggest the samples are highly admixed (Figure 13). Twenty-eight samples (21.4% of the PR unknown samples) show an inferred ancestry ranging from 1.0 to 0.91.

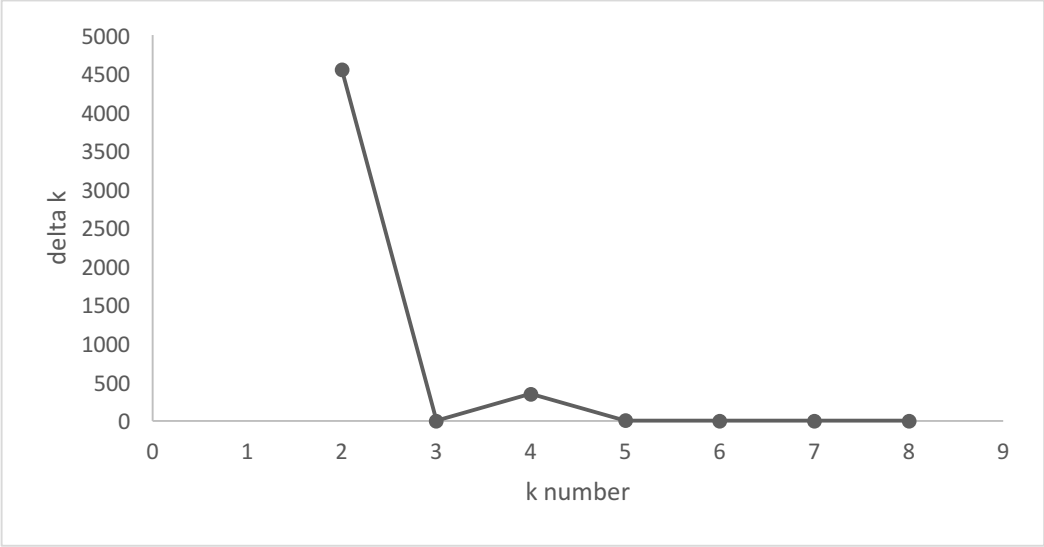


Figure 10. Delta k value using the delta k method of Evanno (2005) for the 154 samples.

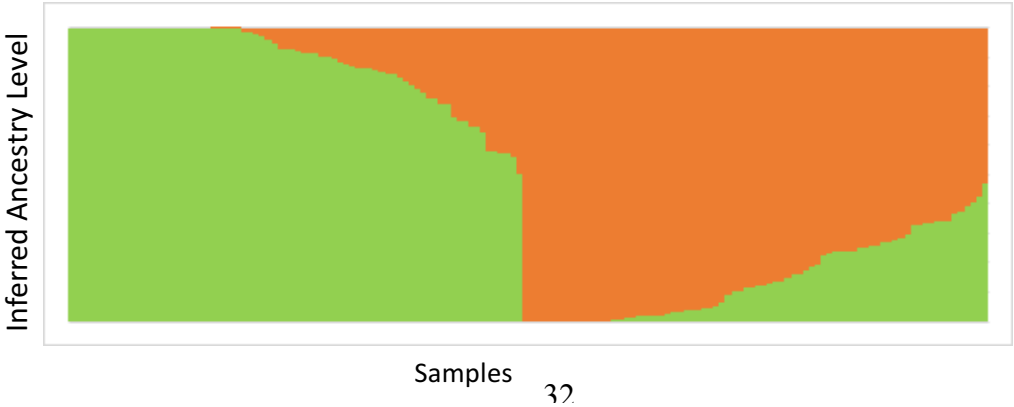


Figure 11. STRUCTURE barplot of the entire 154 sample. The number of groups (k) were identified by the Evanno (2005) delta k method. Both of the identified clusters contains samples from all of the assessed groups. Group 1 is shown in green and group 2 in orange.

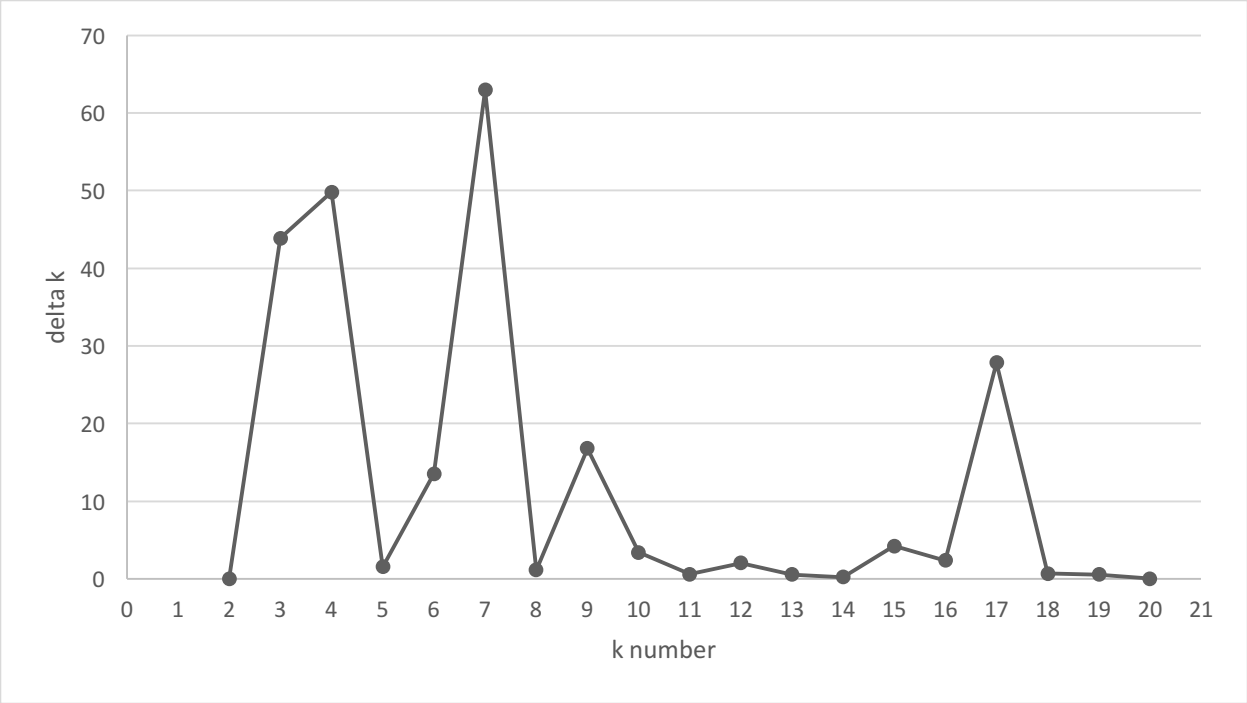


Figure 12. Delta k value using the delta k method of Evanno (2005) for only the Puerto Rico unknown Samples.

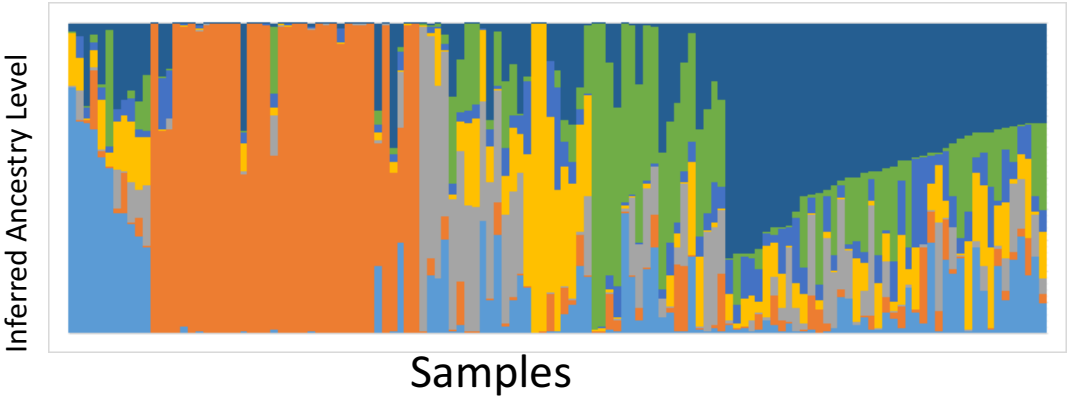


Figure 13. Puerto Rico unknown Samples Structure barplot.

Discussion

The genetic diversity estimators for this analysis revealed similar results as our study that evaluated the genetic diversity of papaya in Puerto Rico using SSR markers. The observed heterozygosity (H_o) for the GBS analysis was 0.226, consistent with the SSR work where H_o was 0.219. Another similarity within our analyses is the general distribution of the samples in the UPGMA dendograms. We observed a similar pattern especially the position of the samples "k14", "Bella", "Tainung No. 1", and "Mona". Likewise, results were reported in Gürcan *et al.* (2016) study in which the genetic diversity of apricots was evaluated with both SSR markers and SNP identification by GBS. In this study 18 SSR markers and 1,162 SNPs were used to construct UPGMA dendograms that have a similar sample distribution but a better resolution is observed on the SNP-based dendogram (Gürcan *et al*., 2016). In our SNP based dendogram, we observe more resolution as well by positioning the commercial samples "Bella" and "Tainung No. 1" closer to the Puerto Rico Unknown samples, thus validating our previous correlation between high genetic diversity and multiple papaya introductions to Puerto Rico. Also, the sample from Mona island was positioned as an out group, again, sustaining our previous hypothesis.

We did not obtain consistent results upon the calculation of the Inbreeding coefficient. In our SSR analysis, we recorded a high F-Index of 0.576 and in the GBS analysis we obtained a low Gis value of -0.092. These results are similar to the ones obtained by OCampo *et al*. (2006) where F-index are low among papaya populations from Venezuela and other Lesser Antilles islands (OCampo *et al.*, 2006). Nevertheless, the population structure analysis of the Puerto Rico unknown samples reveals a highly admixed population which is consistent with the recorded Gis coefficients. The difference between our analyses could be explained by the nature of the molecular markers used. Although both methods are effectively used, a difference in the estimators is frequently observed and discussed (Filippi *et al*., 2015; Emanuelli *et al.*, 2013; Li *et al*., 2010; Hamblin *et al*., 2007). Also, when analyzing the SNP data set, we eliminated the SNPs in LD, a step we cannot perform when using SSR markers. Therefore, if an SSR marker yielded a high homozygosis and it is in LD with another that has also high homozygosis, we could be evaluating redundant data.

Upon the availability of a complete annotated genome, these results will be useful to correlate polymorphisms to phenotype in the commercial and USDA varieties helping towards understanding and development of new varieties. The fact that we obtained similar results in both of our analyses validates the effectivity of both techniques. But, during past 2-3 years GBS is becoming the method of choice because of its high resolution, capacity, time saving method, and economic accessibility. This is the first

ever study in which the genetic diversity of papaya is evaluated using a next generation sequencing tool and may lead to further developments in the field of genomics of papaya.

Conclusions

Conclusions

- This is the first assessment of the genetic diversity of *Carica papaya* using SNPs as molecular markers and GBS as a technique to identify them.
- A low observed heterozygosity (H_o) and Gis were recorded.
- The distribution of samples in the UPGMA dendogram is similar to our other analysis, sustaining our previous hypotheses.
- Two groups were identified upon the population structure analysis for all the samples; for the Puerto Rico Unknown samples, 7 groups were identified.

Recommendations

- The analysis should be repeated upon de availability of a complete annotated genome.
- To include more commercial and USDA germplasm varieties.
- To have the possibility of screening for a certain phenotype, the sample's seeds should be collected and stored.

Appendixes

Appendix A- List of Evaluated Samples

Appendix Table 1. List of evaluated samples for SSR and GBS analyses.

NW- North West, NE- North East, C-Center, SW-South West, USDA- USDA germplasm repository samples, Commercial- commercial varieties.

Number	**Provenance**	Area
1	Aguadilla	NW
3	Aguadilla	NW
11†	Aguadilla	NW
14	Aguadilla	NW
18	Arecibo	NW
19†	Manatí	NW
29†	Aguadilla	NW
33	Aguada	NW
36	Aguada	NW
38†	Isabela	NW
45†	Moca	NW
48	Barceloneta	NW
49†	Camuy	NW
50	Quebradillas	NW
55†	Rincon	NW
56†	Aguada	NW
60	Arecibo	NW
65	Manati	NW
66	Manati	NW
70	Hatillo	NW
73	Arecibo	NW
78	San Sebastian	NW
82	Aguadilla	NW
85	Camuy	NW
86	Moca	NW
89	Aguada	NW
91	San Sebastian	NW
94	Aguadilla	NW
96	Florida	NW
97	Arecibo	NW
101	Aguadilla	NW
102	Aguada	NW
104	Aguada	NW

Cont. Appendix Table 1. List of evaluated samples fro SSR and GBS analyses.

Number	**Provenance**	Area
105	Barceloneta	NW
106	Barceloneta	NW
107	Florida	NW
108	Florida	NW
6	Rio Grande	NE
8†	Carolina	NE
9	Guaynabo	NE
16	Rio Piedras	NE
17	Vega Alta	NE
20	Vega Baja	NE
21	Vega Baja	NE
27	Dorado	NE
30†	Dorado	NE
42	Carolina	NE
51	Vega Baja	NE
53	Rio Piedras	NE
72	Rio Piedras	NE
90	Trujillo Alto	NE
113	Vega Baja	NE
114	Vega Baja	NE
127	Trujillo Alto	NE
128	Toa Baja	NE
134	Vega Baja	NE
137	Naguabo	NE
146	Toa Alta	NE
154	Toa Baja	NE
2	Arroyo	SE
35	Aibonito	SE
37	Salinas	SE
68	Las Piedras	SE
79	Cayey	SE
88	Coamo	SE
116	Coamo	SE
118	Guayama	SE
119	Salinas	SE
120	Santa Isabel	SE
121	Guayama	SE
122	Arroyo	SE
124	Maunabo	SE
125	Yabucoa	SE

Cont. Appendix Table 1. List of evaluated samples fro SSR and GBS analyses.

Number	**Provenance**	Area
130	Vieques	SE
136	Yabucoa	SE
139	San Lorenzo	SE
142	Coamo	SE
143	Salinas	SE
148	Cidra	SE
150	San Lorenzo	SE
151	Cidra	SE
152	Aibonito	SE
153	Salinas	SE
23	Utuado	C
26	Lares	C
58	Naranjito	C
64	Caguas	C
69	Ciales	C
93	Lares	C
95	Morovis	C
109	Ciales	C
110	Ciales	C
111	Morovis	C
112	Morovis	C
115	Jayuya	C
117	Barranquitas	C
126	Caguas	C
129	Corozal	C
131	Lares	C
132	Adjuntas	C
133	Morovis	C
138	Gurabo	C
140	Lares	C
141	Adjuntas	C
145	Ciales	C
147	Utuado	C
149	Utuado	C
13†	Mayaguez	SW
15	Añasco	SW
22	Hormigueros	SW
24†	Peñuelas	SW
25†	Mayaguez	SW
34†	Yauco	SW

Cont. Appendix Table 1. List of evaluated samples fro SSR and GBS analyses.

Number	**Provenance**	Area
39	Mayaguez	SW
41	Yauco	SW
47	Mayaguez	SW
52	Cabo Rojo	SW
54†	Ponce	SW
57	Mayaguez	SW
59	Guayanilla	SW
62	Hormigueros	SW
63	Cabo Rojo	SW
67	Mayaguez	SW
71	Cabo Rojo	SW
74	Mayaguez	SW
75	Cabo Rojo	SW
76	Juana Diaz	SW
77	Mayaguez	SW
80	Yauco	SW
81	Mayaguez	SW
83	Las Marias	SW
84	Mayaguez	SW
87	Las Marias	SW
92	San German	SW
98	Guanica	SW
99	Yauco	SW
100	Las Marias	SW
103	Sabana Grande	SW
Mona	Mona	SW
k14	Panama (Brash, Carica papaya)	USDA
k17	Northern Mariana (Saipan Red, Carica papaya)	USDA
k20	Thailand (Khag Naun, Carica papaya)	USDA
k164	United States (Hawaii, Carica papaya)	USDA
k207	Taiwan (Tainung No. 5, Carica papaya)	USDA
k217	Puerto Rico	USDA
k309	? (Kaek Dum)	USDA
k313	Maradol	USDA
Tainung No.1	-	USDA
Gold Maradol	-	USDA
Red Lady	-	USDA
Known You	-	USDA
Solo	-	USDA
Tropical Red	-	Commercial

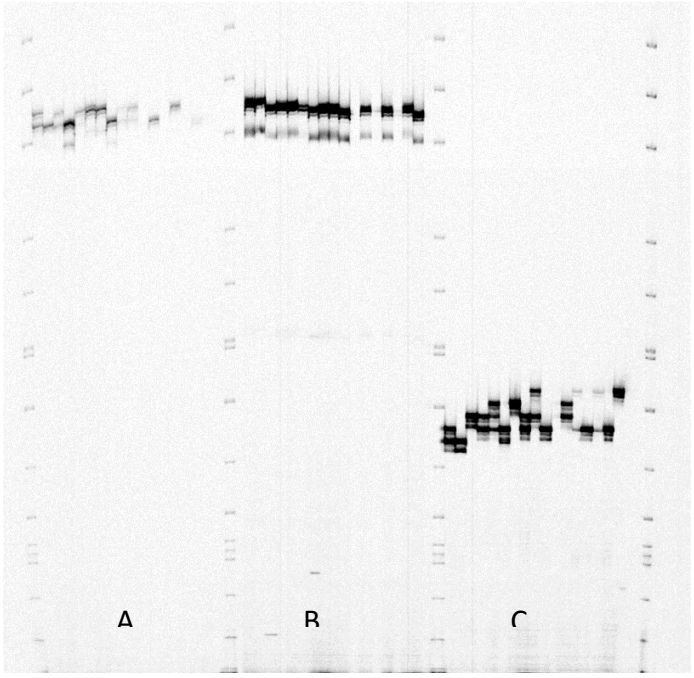
Cont. Appendix Table 1. List of evaluated samples fro SSR and GBS analyses.

Number	**Provenance**	Area
Red maradol	-	Commercial
Known You J†	-	Commercial
Red Lady J†	-	Commercial
Red Matador	-	Commercial
Sun Gold	-	Commercial
Waimanulu X-77	-	Commercial
Red Queen	-	Commercial
Bella	-	Commercial
Taiwan Solo Sunrise	-	Commercial

*- sample was not considered for the SSR analysis

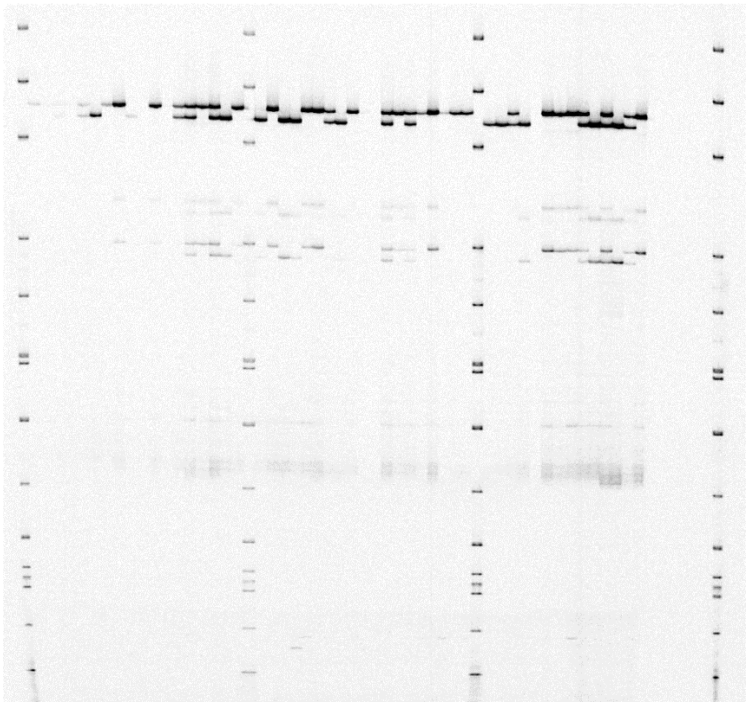
†-sample was not considered for the GBS analysis

Appendix B– SSR Polyacrylamide Gel Images



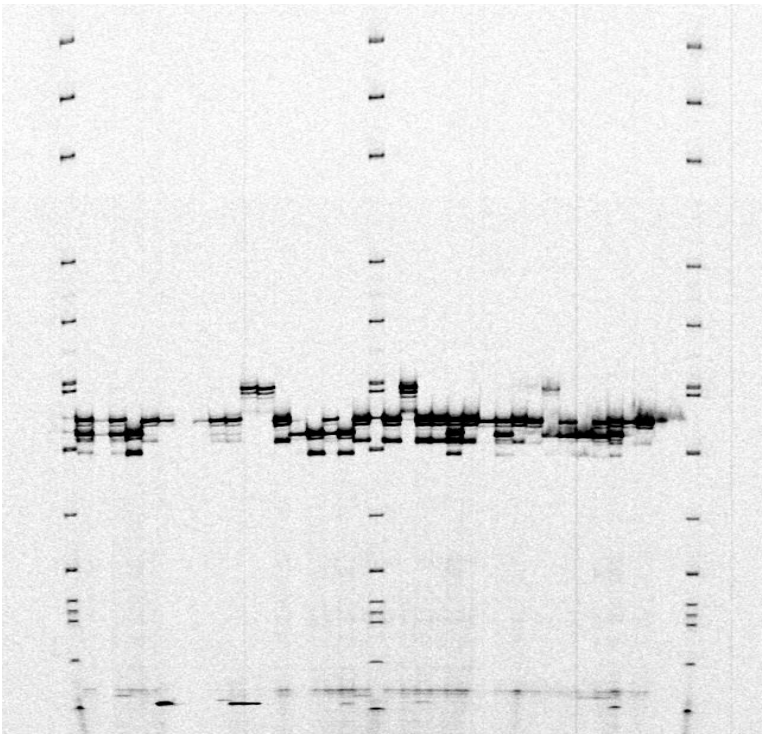
Appendix Figure 1. SSR Polyacrylamide Gel Electrophoresis Image 1

SSR markers a) AJ810495, b) AJ810505, c) CP21 with commercially acquired samples



Appendix Figure 2. SSR Polyacrylamide Gel Electrophoresis Image 2

SSR marker SP7 with Puerto Rico samples from the NE and SE.



Appendix Figure 3. SSR Polyacrylamide Gel Electrophoresis Image 3

SSR marker AJ810491 with Puerto Rico samples from the NW area.

Appendix C- List of SNPs in Linkage Disequilibrium (LD)

Appendix Table 2. Linkage Disequilibrium (LD) SNP Count.

R^2BinMin	R^2BinMax	SNP Count
0	0.01	26485
0.01	0.02	11973
0.02	0.03	7264
0.03	0.04	5319
0.04	0.05	3800
0.05	0.06	3193
0.06	0.07	2423
0.07	0.08	2102
0.08	0.09	1783
0.09	0.1	1458
0.1	0.11	1303
0.11	0.12	1154
0.12	0.13	1055
0.13	0.14	786
0.14	0.15	779
0.15	0.16	674
0.16	0.17	632
0.17	0.18	502
0.18	0.19	461
0.19	0.2	406
0.2	0.21	536
0.21	0.22	433
0.22	0.23	392
0.23	0.24	319
0.24	0.25	276
0.25	0.26	271
0.26	0.27	216
0.27	0.28	202
0.28	0.29	197
0.29	0.3	193
0.3	0.31	184
0.31	0.32	165
0.32	0.33	139
0.33	0.34	165
0.34	0.35	158

Cont. Appendix Table 2. Linkage Disequilibrium (LD) SNP count.

R^2BinMin	R^2BinMax	SNP Count
0.35	0.36	130
0.36	0.37	147
0.37	0.38	137
0.38	0.39	144
0.39	0.4	97
0.4	0.41	121
0.41	0.42	104
0.42	0.43	90
0.43	0.44	102
0.44	0.45	104
0.45	0.46	154
0.46	0.47	110
0.47	0.48	125
0.48	0.49	119
0.49	0.5	39
0.5	0.51	83
0.51	0.52	70
0.52	0.53	64
0.53	0.54	65
0.54	0.55	132
0.55	0.56	74
0.56	0.57	63
0.57	0.58	92
0.58	0.59	65
0.59	0.6	40
0.6	0.61	48
0.61	0.62	27
0.62	0.63	58
0.63	0.64	100
0.64	0.65	157
0.65	0.66	124
0.66	0.67	51
0.67	0.68	36
0.68	0.69	51
0.69	0.7	37
0.7	0.71	34
0.71	0.72	52
0.72	0.73	81

Cont. Appendix Table 2. Linkage Disequilibrium (LD) SNP count.

R^2BinMin	R^2BinMax	SNP Count
0.73	0.74	112
0.74	0.75	48
0.75	0.76	38
0.76	0.77	42
0.77	0.78	45
0.78	0.79	66
0.79	0.8	56
0.8	0.81	64
0.81	0.82	125
0.82	0.83	29
0.83	0.84	34
0.84	0.85	29
0.85	0.86	23
0.86	0.87	48
0.87	0.88	14
0.88	0.89	59
0.89	0.9	48
0.9	0.91	56
0.91	0.92	29
0.92	0.93	18
0.93	0.94	25
0.94	0.95	31
0.95	0.96	25
0.96	0.97	0
0.97	0.98	0
0.98	0.99	0
0.99	1	2812
NaN	NaN	313774

References

Agarwal, M., Shrivastava, N., & Padh, H. 2008. Advances in molecular marker techniques and their applications in plant sciences. Plant cell reports. 27(4):617- 631.

Aikpokpodion, P.O. 2012. Assessment of genetic diversity in horticultural and morphological traits among papaya (*Carica papaya*) accessions in Nigeria. Fruits. 67:173–187

Alonso, M., Alor, B., García, O., Moreno, Q., Teyer, S., and Felipe, L. 2009. Caracterización de accesiones de papaya (Carica papaya L.) a través de marcadores AFLP en Cuba Characterising Cuban papaya accessions (Carica papaya L.) by AFLP markers. Rev. Colomb. Biotecnol. 2:31–39

Antrop, M. 2004. Landscape change and the urbanization process in Europe. Landsc. Urban Plan. 67:9–26

Arumuganathan, K., and Earle, E. D. 1991. Nuclear DNA Content of Some Important Plant Species. Plant Mol. Biol. Report. 9:208–218

Asudi G., Ombwara F.K., Rimberia F.K., Nyende A.B., Ateka EM, L.S. Wamocho. 2013. Evaluating diversity among kenyan papaya germplasm using simple sequence repeat markers. African J. Food, Agric. Nutr. Dev. 13:7307–7324

Asudi GO, Ombwara FK, Rimberia FK, Nyende AB, Ateka EM. 2011. Morphological Characterization of Kenyan Papaya Germplasm (Carica papaya L .). Acta Hortic 383–390

Badillo VM. 1971. Monografía de la familia *Caricaceae* e. Publicado por la Asociación de profesores.Universidad Central de Venezuela, Maracay.

Badillo VM. 1993. *Caricaceae* e. Segundo Esquema. Rev Fac Agron Univ Centr Venezuela 43:1–111

Badillo VM .2000. Carica vs Vasconcella St.-Hil. (*Caricaceae*) con la rehabilitación de este último. Ernstia 10:74–79

Bateson MF, Lines RE, Revill P, Chaleeprom W, Ha C V, Gibbs AJ, Dale JL. 2002. On the evolution and molecular epidemiology of the potyvirus Papaya ringspot virus. J Gen Virol 83: 2575–2585

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Britton N.L. 1915. The Vegetation of Mona Island. *Missouri Botanical Garden Press*, *2*:1, 33–58.

Becker S. 1958. The Production of Papain--An Agricultural Industry for Tropical America. Econ Bot 12: 62–79

Brown, J. E., Bauman, J. M., Lawrie, J. F., Rocha, O. J., and Moore, R. C. 2012. The Structure of Morphological and Genetic Diversity in Natural Populations of Carica papaya (Caricaceae) in Costa Rica. Biotropica. 44:179–188

Carro-Figueroa, V. 2002. Agricultural Decline and Food Import Dependency in Puerto Rico: A Historical Perspective on the Outcomes of Postwar Farm and Food Policies. Carib. Studies *30*(2): 77–107.

Carvalho FA, Renner SS (2012) A dated phylogeny of the papaya family (Caricaceae) reveals the crop's closest relatives and the family's biogeographic history. Mol Phylogenet Evol 65: 46–53

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 4:7

Chen C, Yu Q, Hou S, Li Y, Eustice M, Skelton RL, Veatch O, Herdes RE, Diebold L, Saw J, *et al.* 2007. Construction of a Sequence-Tagged High-Density Genetic Map of Papaya for Comparative Structural and Evolutionary Genomics in Brassicales. Genetics 177: 2481–2491

da Silva F, Pereira M, Junior P, Pereira T, Viana A, Daher R, Ramos H, Ferreguetti G. 2007. Evaluation of the sexual expression in a segregating BC 1 papaya population. Crop Breed Appl Biotechnol 7(1):16–23

De Oliveira EJ, dos Santos Silva A, de Carvalho FM, dos Santos LF, Costa JL, de Oliveira Amorim VB, Dantas JLL .2010. Polymorphic microsatellite marker set for Carica papaya L. and its use in molecular-assisted selection. Euphytica 173: 279–287

de Oliveira, E. J., Amorim, V. B. O., Matos, E. L. S., Costa, J. L., Silva Castellen, M., Pádua, J. G., and Dantas, J. L. L. 2010. Polymorphism of Microsatellite Markers in Papaya (Carica papaya L.). Plant Mol. Biol. Report. 28:519–530

de Oliveira, J. G., and Vitória, A. P. 2011. Papaya: Nutritional and pharmacological characterization, and quality loss due to physiological disorders. An overview. Food Res. Int. 44:1306–1313

Departamento de Agricultura de Puerto Rico. 2009. Distribución del valor de la producción agrícola 2009/10 en orden de importancia económica. 98–99

Doyle, J. 1991. CTAB Total DNA Isolation. Mol. Tech. Taxon. 57:283–293

Duany, J. 2005. Dominican migration to Puerto Rico: A transnational perspective. Cent. J. 17:243–268

Earl, D. A., & VonHoldt, B. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Gen Res, *4*, 359–361.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5: 435–45

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PloS One*, *6*:5, 1–10.

Emanuelli F, Lorenzi S, Greskowiak L, Catalano V, Stefanini M, Troggio M, Myles S, Martinez-Zapater JM, Zyprian E, Moreira FM, Grando SM. 2013. Genetic Diversity and population structure assessed by SSR and SNP markers in a large gremplasm collection of grape. BMC Plant Bio. 13:39

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

Filippi CV, Aguirre N, , Zubrzycki J, Puebla A, Cordes D, Moreno MV, Fusari CM, AlvarezD, Heinz RA, Hopp HE, Paniego NB, Lia VV. 2015. Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. BMC Plant Bio. 15: 52

Food and Agriculture Organization of the United Nations. 2010. The Second Report on The State of the World's Plant Genetic Resources for Food and Agriculture. pp 1–16.

Food and Agriculture Organization of the United Nations.2015. FAOSTAT database. Available at http://faostat.fao.org/ (checked on 10/12/16)

Fuentes, G., and Santamaria, J. M. 2014. Papaya (*Carica papaya* L.): Origin, Domestication, and Production. Pages 3–15 in: Genetics and Genomics of Papaya, Plant Genetics and Genomics: Crops and Models 10, R. Ming and P.H. Moore, eds. Springer Science+Business Media, New York, NY.

Ganal MW, Altmann T, Röder MS. 2009. SNP Identification in crop plants. Curr Opinion in Plant Bio. 12:211-217

Goenaga R, Irizarry H, Rivera-Amador E. 2001. Yield and fruit quality of papaya cultivars grown at two locations in Puerto Rico. J Agric Univ Puerto Rico 85: 127–134

Gonsalves D (1998) Control of papaya ringspot virus in papaya: a case study. Annu Rev Phylopathology 36: 415–437

Gürcan K, Teber S, Ercilsli S, Ugurtan Y. 2016. Genotyping by Sequencing (GBS) in Apricots and Genetic Diversity Assessment with GBS-Derived Single-Nucleotide Polymorphisms (SNPs). Biochem Genet. 54(6):854-885

Hamblin MT, Warburton ML, Buckler ES. 2007. Empirical Comparison of Simple Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize Diversity and Relatedness. PLoS ONE 2(12): e1367

Hardisson A, Rubio C, Baez A, Martin MM, Alvarez R (2001) Mineral composition of the papaya (Carica *papaya* variety Sunrise) from Tenerife Island. Eur Food Res Technol 212:175–181

Hoshino AA, Bravo JP, Nobile PM, Morelli KA (2012) Microsatellites as Tools for Genetic Diversity Analysis. *In* M Caliskan, ed, Genet. Divers. Microorg., 1st ed. InTech, p 382

Idrees, M., and Irshad, M. 2015. Molecular Markers in Plants for Analysis of Genetic Diversity: A Review. Eur. Acad. Res. 2:1513–1540

Ivanov KI, Eskelin K, Lohmus A, Makinen K. 2014. Molecular and cellular mechanisms underlying potyvirus infection. J Gen Virol 95: 1415–1429

Jakobsson, M., and Rosenberg, N. A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. 23:1801–1806

Jarne P, Lagoda JL. 1996. Microsatellites, from molecules to populations and back. Trends Ecol Evol 11: 424–429

Jiménez D. 2002. Caracterización y estudio de la diversidad genética de los géneros *Vasconcellea* y *Carica* (Caricaceae) en Colombia y Ecuador con marcadores isoenzimáticos. Dissertation. Universidad de Caldas, Manizales, Colombia

Jimenez, V. M., Mora-Newcomer, E., and Gutierrez-Soto, M. V. 2014. Biology of the Papaya Plant. Pages 17–33 in: Genetics and Genomics of Papaya, Plant Genetics and Genomics: Crops and Models 10, R. Ming and P.H. Moore, eds. Springer Science+Business Media, New York, NY.

Jones ES, Sullivan H, Bhattramakki D, Smith JSC. 2007 A comparison of simple sequence repeats and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (Zea mays L.). Theor Appl Genet. 115:361–71.

Junta de Planificacion Puerto Rico. 2016. External Trade Data. Available at: http://www.jp.gobierno.pr/. (checked on 01/25/2017)

Kesawat, M. S., & Das, B. K. 2009. Molecular markers: It's application in crop improvement. *Journal of Crop Science and Biotechnology*, *12*(4), 169-181.

Kim MS, Moore PH, Zee F, Fitch MMM, Steiger DL, Manshardt RM, Paull RE, Drew RA, Sekioka T, Ming R. 2002. Genetic diversity of Carica papaya as revealed by AFLP markers. Genome 45: 503–512

Letunic, I., and Bork, P. 2007. Interactive Tree of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics. 23:127–128

Li Wang M, Barkley NA, Jenkins TM. 2009. Microsatellite Markers in Plants and Insects. Part I: Applications of Biotechnology. Genes, Genomes and Genomics 3: 54–67

Li Y-H, Li W, Zhang C, Yang L, Chang R-Z, Gaut BS, Qiu L. 2010. Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytol. 188:242–53.

Liu, J., Muse, S., and Bruce, W. 2005. Power Marker V. 3.25: Integrated Analysis Environment for genetic marker data. Bioinformatics. 21:2128-2129

Madarbokus, S., & Ranghoo-Sanmukhiya, V. M. 2012. Identification of genetic diversity among papaya varieties in Mauritius using morphological and molecular markers. *Int. J. Life Sci. Bt. and Pharm. Res*, 1(4), 153-163.

Matos, E. L. S., Oliveira, E. J., Jesus, O. N., and Dantas, J. L. L. 2013. Microsatellite markers of genetic diversity and population structure of Carica papaya. Ann. Appl. Biol. 163:298–310

Meirmans PG, van Tienderen PH. 2004. genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. Mol. Ecology Res. 4(4):792-794

Ming R. 2014. Genetics and Genomics of Papaya. Springer Science+Business Media, New York,NY

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly B V, Lewis KLT, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452: 991–6

Ming, R., Yu, Q., and Moore, P. H. 2007. Sex determination in papaya. Semin. Cell Dev. Biol. 18:401–408

Morgante M, Olivieri AM. 1993. PCR-amplified microsatellites as markers in plant genetics. Plant Journal 3, 175-182

Morton, J. 1987. Papaya. p. 336–346. In: Fruits of warm climates. Julia F. Morton, Miami, FL.

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. 2012. SNP Markers and their implact in plant breeding. Int. Jour of Plant Genomics. 2012:1-11

Muller,F., Voccia, M., Ba, A., Bouvet J-M. 2009. Genetic diversity and gene flow in a Caribbean tree Pterocarpus officinalis Jacq .: a study based on chloroplast and nuclear microsatellites. *Genetica*, 135: 185–198.

Nagarajan N, Navajas-Pérez R, Pop M, Alam M, Ming R, Paterson AH, Salzberg SL. 2008. Genome-Wide Analysis of Repetitive Elements in Papaya. Trop Plant Biol 1: 191–201

Niggemann, M., Jetzkowitz, J., Brunzel, S., Wichmann, M. C., and Bialozyt, R. 2009. Distribution patterns of plants explained by human movement behaviour. Ecol. Modell. 220:1339–1346

Ocampo Perez, J. 2007. Papaya genetic diversity assessed with microsatellite markers in germplasm from the caribbean region. Acta Hortic. 740:93–102

Ocampo Perez, J., D'Eeckenbrugge, G. C., Bruyere, S., Bellaire, L. de L., and Ollitrault, P. 2006. Organization of Morphological and genetic diversity of Caribbean and Venezuelan papaya germplasm. Fruits. 61:25–37

Ocampo Perez, J., Dambier, D., Ollitrault, P., Coppens d'Eeckenbrugge, G., Brottier, P., Froelicher, Y., and Risterucci, A.-M. 2006. Microsatellite markers in Carica papaya L.: isolation, characterization and transferability to Vasconcellea species. Mol. Ecol. Notes. 6:212–217

Ocampo, J. A., Dambier, D., Ollitrault, P., Coppens d'Eeckenbrugge, G., Brottier, P., Risterucci, A-M. 2004. Development of Microsatellite Markers in Papaya: Isolation , Characterization and Cross Amplification in Mountain Papayas. Proc. Interamer. Soc. Trop. Hort. 48:90–93

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Lúcia M, Vieira C. 2006. Origin, evolution and genome distribution of microsatellites. Genet Mol Biol 29: 294–307

Paterson AH, Felker P, Hubbell SP, Ming R. 2008. The Fruits of Tropical Plant Genomics. Trop Plant Biol 1: 3–19

Peakall, R., and Smouse, P. E. 2012. GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics. 28:2537–2539

Peakall, R., Gilmore, S., Keys, W., Morgante, M., & Rafalski, A. 1998. Cross-species amplification of soybean (Glycine max) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Molecular biology and evolution, 15(10): 1275-1287.

Powell, W., Morgante, M., Andre, C.,Hanafey, M., Vogel, J., Tingey, S., & Rafalski, A. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular breeding, 2(3):225-238.

Prest RL. 1955. Unfruitfulness in papaya. Qd Agric J 81:144–148

Rabelo da Costa, F., Nair, T., Pereira, S., Paula, A., Gabriel, C., and Pereira, M. G. 2011. ISSR markers for genetic relationships in Caricaceae and sex differentiation in papaya. Crop Breeding and Applied Biotechnology.11:352-357

Rafalski A. 2002. Applications of Single Nucleotide Polymorphisms in crop genetics. Curr Opinion in Plant Bio. 5:94-100

Rieger, J. E. 2009. Genetic and morphological diversity of natural populations of *Carica papaya*. Dissertation. Miami University, Oxford, Ohio, United States of America.

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

Santos SC, Ruggiero C, Lacerda C, Petrarolha S, Lemos GM. 2003. A microsatellite library for *Carica papaya* L. CV. Sunrise Solo. Rev Bras Frutic 25: 263–267

Schlotterer C, Harr B. 2001. Microsatellite Instability. Encycl Life Sci 1–4

Schlotterer, C. 2004. The evolution of molecular markers—just a matter of fashion? Nature Reviews Genetics, *5*(1):63-69.

Sengupta, S., Das, B., Prasad, M., Acharyya, P., and Ghose, T. K. 2013. A comparative survey of genetic diversity among a set of Caricaceae accessions using microsatellite markers. Springerplus. 2:345-355

Setrini G. 2012. Cultivating New Development Paths: food and agriculture entrepreneurship in Puerto Rico Gustavo Setrini Department of Political Science Massachusetts Institute Technology.

Siar S V., Beligan G a., Sajise a. JC, Villegas VN, Drew RA. 2011. Papaya ringspot virus resistance in *Carica papaya* via introgression from *Vasconcellea quercifolia*. Euphytica 181: 159–168.

Somasundaram, S. T., & Kalaiselvam, M. 2011. Molecular tools for assessing genetic diversity. International Training Course on Mangroves and Biodiversity, Annamalai University, India, 82-91.

Sudha, R., Singh, D. R., Sankaran, M., Singh, S., Damodaran, V., and Simachalam, P. 2013. Genetic diversity analysis of papaya (Carica papaya L.) genotypes in Andaman Islands using morphological and molecular markers. 8:5187–5192

Teixeira, J. A., Rashid, Z., Tan, D., Dharini, N., Gera, A., Teixeira, M., Jr, S., and Tennant, P. F. 2007. Papaya (Carica papaya L .) Biology and Biotechnology. Tree and Forestry Science and Biotechnology, Global Science Books (Lond). pp47–73

Uitdewilligen JGAML, Wolters AA, D'hoop BB, Borm TJA, Visser RGF, van Eck H. 2013. A Next-Generation Sequencing Method for Genotyping by-Sequencing of Highly Heterozygous Autotetraploid Potato. PLoS ONE 8(5): e62355.

United States Coast Guard.2016. Alien Migrant Interdiction Statistics. Available at: https://www.uscg.mil/hq/cg5/cg531/AMIO/amio.asp. (checked on 10/13/2016)

Van Droogenbroeck B, Breyne P, Goetghebeur P, Romeijn-Peeters E, Kyndt T, Gheysen G. 2002. AFLP analysis of genetic relationships among papaya and its wild relatives (Caricaceae) from Ecuador. Theor Appl Genet 105: 289–297

Vanburen, R., Zeng, F., Chen, C., Zhang, J., Wai, C. M., Han, J., Aryal, R., Gschwend, A. R., Wang, J., Na, J., Huang, L., Zhang, L., Miao, W., Gou, J., Arro, J., Guyot, R., Moore, R. C., Wang, M., Zee, F., Charlesworth, D., Moore, P. H., Yu, Q., and Ming, R. 2015. Origin and domestication of papaya Y h chromosome. Genome Res. 25: 1-10

Varshney RK, Graner A, Sorrells ME. 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23: 48–55

Vidal, N. M., Grazziotin, A. L., Ramos, H. C. C., Pereira, M. G., and Venancio, T. M. 2014. Development of a Gene-Centered SSR Atlas as a Resource for Papaya (Carica papaya) Marker-Assisted Selection and Population Genetic Studies C. Chen, ed. PLoS One. 9:e112654

Wadsworth, F. H. 1972. The Historical Resources of Mona Island. Las Islas de Mona y Monito: Una evaluación de sus recursos naturales e históricos. San Juan: Junta de Calidad Ambiental. [2 volumes.]

Wang, M.L., Barkley, N. A., and Jenkins, T. M. 2009. Microsatellite Markers in Plants and Insects. Part I: Applications of Biotechnology. Genes, Genomes and Genomics. 3:54–67

Weingartner, L. A., & Moore, R. C. 2012. Contrasting patterns of X/Y polymorphism distinguish Carica papaya from other sex chromosome systems.*Molecular biology and evolution*, *29*(12), 3909-3920.

Wendel, J. F., Brubaker, C. L., & Percival, A. E. 2017. Genetic Diversity in *Gossypium hirsutum* and the Origin of Upland Cotton. American Journal of Botany, *79*: 1291–1310.

Zambrana-Echevarría, C., de Jesús-Kim, L., Márquez-Karry, R., Siritunga, D., and Jenkins, D. 2016. Diversity of Papaya ringspot virus isolates in Puerto Rico. HortScience. 51:362–369