University of São Paulo
"Luiz de Queiroz" College of Agriculture

Genetic structure, mating system and domestication of annatto (*Bixa orellana* L.) using molecular markers

Gabriel Dequigiovanni

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2017

Gabriel Dequigiovanni
Bioprocess and Biotechnology Engineering

Genetic structure, mating system and domestication of annatto (*Bixa orellana* L.) using molecular markers

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. ELIZABETH ANN VEASEY

Thesis presented to obtain the degree of Doctor in Science. Area: Genetics and Plant Breeding

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Dequigiovanni, Gabriel

Genetic structure, mating system and domestication of annatto (*Bixa Orellana* L.) using molecular markers / Gabriel Dequigiovanni. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2017.

127 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. *Bixa orellana* 2. Microssatélites 3. Amazônia 4. Diversidade genética 5. Estrutura genética 6. Domesticação 7. Cloroplasto I. Título

To those who I love without measures or proportions,
I just love them with all my heart.
My parents Alda and Enio,
My brother Diego,
My wife Patricia.

ACKNOWLEDGEMENTS

To the University of São Paulo, Superior School of Agriculture "Luiz de Queiroz" – ESALQ, and the Genetics and Plant Breeding graduate program for the opportunity to obtain the PhD degree.

To the São Paulo Research Foundation (FAPESP) for the financial support: grants 2013/08884-5 and 2016/05912-6, that were fundamental for the accomplishment of this work.

To the National Council for Scientific and Technological Development (CNPq) for the scholarship and financial support.

To Dr. Elizabeth Ann Veasey, my advisor and leader of the Plant Ecological Genetics Laboratory, for the confidence, all opportunities, discussions and the great help in the writing of this work and related projects. For never measure efforts in accomplishing this project, for the excellent corrections, and for the total support in all my decisions. I also thank you for being always on my side when the difficulties arose. To you Betty, my most sincere thanks.

To my dear parents, Enio and Alda, for the values, principles and unconditional love. Without the inspiration, guidance, and support that you have given me, I might not be the person I am today. I certainly owe everything to you.

To my brother Diego and my sister-in-law Carla, for the friendship, and even far away always sent many messages of encouragement.

To Schaker family, Jamil, Soili, Felipe and Leticia, for the support during all these years.

To Dr. Charles Roland Clement (INPA) for providing samples, and all the assistance in the delineation of the studies, in the corrections, in the articles writing, and especially for all the opportunities.

To Dr. Eliane Gomes Fabri (IAC) for helping in the selection and supply of plant materials, as well as the availability to solve doubts, and all her efforts that allowed me to carry out this work.

To my friends and colleagues of the Plant Ecological Genetics Laboratory: Jucelene, Mariana, Sueme, Nancy, Santiago, Adriana, Tatiane, Thais, Flaviane, Iris, Ana Flavia, Cattleya, Paula, Marcela, for the friendship and support during the development of this project.

To all the professors from the Genetics Department - ESALQ/USP, for the great teachings.

To the staff of the Genetics Department - ESALQ/USP, especially to Leia, for all the assistance.

To Marcos Cella (Marcão), technician of the Evolutionary Ecology and Applied Genetics Laboratory, for the help in the accomplishment of the experiments, and friendship.

To my special friends Sueme, Alessandro, Melina and Maisa, for making my days much happier.

To my friends Felipe and Juliana, who provided moments of great joy in this period.

To my friend Alessandro Alves Pereira for the great assistance in the collections and statistical analyzes, besides the discussions that have yielded results in this work and mainly for the friendship.

To my friend Santiago for the great help in the collections, laboratory analyzes, for the assistance in the period that I was in Manaus, and mainly for the friendship.

To Prof. Dr. Maria Teresa Gomes Lopes, for the opportunities, and especially for all assistance during the time I spent in Manaus.

To Prof. Dr. Maria Imaculada Zucchi, for the aid in the development of the enriched library and the great teachings.

To Prof. Dr. Marcos Siqueira, for the friendship and especially for opening the doors of ESALQ.

To Prof. Dr. Paul Gepts and his students for the great reception during my internship period at UCDavis, for the teachings and contributions to this work.

To Dr. Eduardo de Andrade Bressan for the suggestions during the development of the project, in writing and in the laboratory routines.

To Paulo and Marta from ITAL for sharing the phitochemical analyzes of the accessions from the germplasm bank;

To the "Orquestra Educacional de Piracicaba (OEP)" for the great moments we shared. Especially to maestro Ivan and my teachers Christine, Jacqueline and Marina.

To the technicians of EMATER-RO and EMATER-PA for assistance in the collections;

To all those who directly or indirectly contributed to the accomplishment of this work.

Especially to my wife Patricia, for her love, patience, understanding, support, encouragement and affection at all times. Without you nothing of this would be possible, thank you very much for everything.

"If the sun refused to shine, I would still be loving you
When mountains crumble to the sea, there'll still be you and me…
And so today, my world it smiles; your hand in mine we walk the miles
***Thanks to you** it will be done. For you to me are the only one"*

J. P. Page and R. A. Plant (1969)

To God, for the gift of life, for always enlightening and guiding me.

"Mama told me when I was young
Come sit beside me, my "little" son
And listen closely to what I say
And if you do this it will help you some sunny day

Oh, take your time, don't live too fast
Troubles will come, and they will pass
Follow your heart, and nothing else
And be a simple kind of man
be something you love and understand

And don't forget, son there is someone up above…"

Lynyrd Skynyrd (1973)

SUMMARY

RESUMO

Estrutura genética, sistema reprodutivo e domesticação de urucum (*Bixa orellana* L.) utilizando marcadores moleculares

Domesticação de plantas é um processo evolutivo que pode gerar uma série de modificações nas plantas para aumentar a adaptação para o cultivo e utilização pelos humanos. Estas modificações podem diminuir a aptidão das plantas no habitat selvagem, porém, aumentando sua aptidão para exploração humana. Urucum (*Bixa orellana* L.) é uma planta arbustiva domesticada na Amazônia a partir de populações de *Bixa orellana* var. *urucurana*. Esta tese apresenta um entendimento mais aprofundado sobre a domesticação, sistema reprodutivo e diversidade genética e estrutura de urucum e seu ancestral selvagem no Brasil. No primeiro estudo, um novo conjunto de 32 locos microssatélites foram isolados a partir de uma biblioteca genômica enriquecida com microssatélites, dos quais 12 foram polimórficos em populações de urucum selvagem e cultivado. No segundo estudo, a diversidade e estrutura genética de populações selvagens de urucum na Amazônia brasileira foram caracterizadas usando 16 marcadores microssatélites. Elevada estrutura populacional, e correlações positivas entre distancias genéticas e geográficas foram observadas, sugerindo que a diferenciação genética é resultante de isolamento geográfico. Adicionalmente, Modelagem de Nicho Ecológico foi utilizada para caracterizar a distribuição potencial desta variedade no norte da América do Sul e observamos que o Sul de Rondônia, a bacia do rio Madre de Dios, os Llanos de Mojos e de Orinoco e oeste do Equador são áreas de alta probabilidade de ocorrência de urucum selvagem, fornecendo informações importantes para novas amostragens e conservação. No terceiro estudo, 16 locos de microssatélites e quatro compostos fitoquímicos foram utilizados para avaliar a diversidade genética de 63 acessos do banco de germoplasma de urucum do Instituto Agronômico (IAC). Em ambas as análises, houve uma tendência de separação dos acessos de Rondônia, norte do Brasil, dos acessos do Sudeste. Os acessos de Rondônia apresentaram elevados valores para todos os compostos fitoquímicos e também apresentaram altos níveis de diversidade genética. Alguns acessos apresentaram níveis de bixina acima da média e são considerados materiais promissores para uso em programas de melhoramento genético de urucum. No quarto estudo, 12 locos microssatélites foram utilizados para determinar o sistema de cruzamento de uma população de urucum de Rondon do Pará, PA. A taxa de cruzamento multilocos indicou um sistema misto de cruzamento para esta população. A endogamia biparental também contribuiu para a taxa de autofecundação. Cruzamentos entre indivíduos aparentados também foram observados. Devido ao sistema misto, a coleta de sementes de polinização aberta para fins de conservação e melhoramento genético deve incluir pelo menos 60 plantas para assegurar uma amostragem representativa. No quinto estudo, a distribuição da diversidade genética entre amostras de urucum cultivado de quintais de comunidades ribeirinhas dos principais rios da Amazônia Brasileira, além de plantações ao longo das rodovias dos estados do Rondônia e Pará, além do Sudeste do Brasil foi caracterizada. As amostras coletadas apresentaram moderados níveis de diversidade genética e moderados a altos níveis de fluxo gênico entre os grupos geográficos, principalmente devido ao intercambio de semente entre agricultores. Contudo, análises Bayesianas e de agrupamento indicaram uma tendência de agrupamento baseado na origem geográfica das amostras. Isolamento por distância também foi observado de acordo com o teste de Mantel. No último estudo, amostras de urucum selvagem e cultivado da Amazônia brasileira foram comparados utilizando 16 locos microssatélites e duas regiões de DNA cloroplastidial. Uma clara separação entre cultivados e selvagens, suportada por altos valores de F_{ST} em ambas as análises foi observado. Amostras selvagens apresentaram altas taxas de diversidade em relação aos cultivados, parcialmente por não sofrem seleção antrópica como acontece nas variedades cultivadas. Os dados sugerem a existência de relações genéticas entre urucum selvagem e cultivado, indicado por moderados níveis de fluxo gênico. Os resultados também demonstraram a proximidade entre grupos de urucum selvagem e cultivados de Rondônia e da bacia do Rio Madeira. Esta proximidade fornece indícios que a domesticação de urucum iniciou nesta região a partir de *B. orellana* var. *urucurana*.

Palavras-chave: *Bixa orellana*; Microssatélites; Amazônia; Diversidade genética; Estrutura genética; Domesticação; Cloroplasto

ABSTRACT

Genetic structure, mating system and domestication of annatto (*Bixa orellana* L.) populations from Brazil using molecular markers

Plant domestication is an evolutionary process that leads to several modifications in plants to increase adaptation to cultivation and utilization by humans. These modifications may decrease the fitness of plants in the wild habitat but increase it for human exploitation. Annatto (*Bixa orellana* L.) is a shrubby plant domesticated in Amazonia from wild annatto (*Bixa orellana* var. *urucurana*) populations. This thesis presents a more in-depth understanding of the domestication, mating system and genetic diversity and structure of annatto and its wild ancestor in Brazil. In the first study, a new set of 32 microsatellite loci isolated from a microsatellite-enriched genomic library was developed, of which 12 were polymorphic in populations of both cultivated and wild annatto. In the second study, the genetic diversity and structure of wild annatto populations in Brazilian Amazonia were characterized with 16 microsatellite markers. High population structure and positive correlation between genetic and geographic distances were found, suggesting that genetic differentiation might be caused by geographic isolation. Additionally, Ecological Niche Modeling was used to characterize the potential geographical range of this variety in northern South America and detected that South Rondônia, Madre di Dios River basin, Llanos de Mojos, Llanos de Orinoco and eastern Ecuador are highly suitable areas for wild annatto to occur, providing additional targets for future exploration and conservation. In the third study, 16 microsatellite loci and four phytochemical compounds were used to evaluate the genetic diversity of 63 accessions from the annatto germplasm bank at the Agronomic Institute (IAC). In both molecular and phytochemical analysis the results tended to separate the accessions from Rondônia, northern Brazil, from the Southwestern accessions. Rondônia accessions showed higher values for all the phytochemical compounds and higher levels of genetic diversity. Some accessions presented bixin levels well above the average and are promising materials to be used in genetic improvement programs. In the fourth study, 12 microsatellite loci were used to determine the mating system of a cultivated population of annatto from Rondon do Pará, PA. Multilocus outcrossing rate indicated a mixed mating system for this population. Biparental inbreeding also contributed to the selfing rate in this population. Crossings among related individuals were also observed. Due to this mixed breeding system, the collection of open-pollinated seeds for plant breeding or conservation purposes should include at least 60 plants to ensure a representative sample. In the fifth study, the amount and distribution of genetic diversity among samples of cultivated annatto from homegardens of riverside communities along the major rivers in Brazilian Amazonia, and from farmer´s fields along highways, in the States of Rondônia and Pará, and Southeastern Brazil was characterized. The samples collected presented moderate levels of genetic diversity, and moderate to high levels of admixture between geographic groups, occurring mainly due to exchange of seeds among farmers. However, cluster and Bayesian analyses showed a tendency to group samples based on their geographic origin. Isolation by distance was observed, according to Mantel's test. In the last study, wild and cultivated annatto samples from Brazilian Amazonia were compared using 16 microsatellite loci and two cpDNA regions. A clear separation between wild and cultivated annatto, supported by high values of F_{ST} in both analyses was observed. Wild samples presented higher rates of diversity in relation to cultivated, partly because these populations did not suffer anthropic selection, as in the cultivated varieties. The data suggest the existence of genetic relationship between wild and cultivated annatto, indicated by moderate levels of gene flow. The results also showed the proximity between groups of cultivated and wild accessions from Rondônia and the Madeira River basin. This proximity provides indications that annatto started its domestication in this area from *B. orellana* var. *urucurana*.

Keywords: *Bixa Orellana*; SSR; cpDNA; Genetic diversity; Genetic structure; Domestication; Amazonia

1. INTRODUCTION

Domestication of plants and animals is one of the most important processes taking place in the last 13,000 years, becoming a fundamental aspect in the transition from hunting-gathering to agriculture (Badr and El-Shazly, 2012). It was also a prerequisite to the rise of civilizations, and had transformed the global demography (Diamond, 2002). Plant domestication is the result of a selection process that leads to increased adaptation of plants to cultivation and utilization by humans (Brown, 2010). It is a continuing process over a long period of time, leading to significant changes in genotype and phenotype of the plants that are subjected to this process (Gepts, 2004). Plant domestication encompasses many evolutionary modifications that may decrease the fitness of a plant in wild habitat but increase it for human exploitation. The moment in which a crop can be considered domesticated is somewhat speculative. However, the full dependence of a plant on humans for survival is the fullest extent of domestication (Meyer and Purugganan, 2013).

The distinction between wild and cultivated is one of the basic questions of plant domestication (Terrell *et al.*, 2003). Domesticated plants present a number of morphological changes when compared to their wild ancestors. Harlan (1992) called these changes as "Domestication Syndrome". The traits in this syndrome include those increasing adaptation to cultivation and desirability of human consumption and use (Harlan, 1992; Gepts, 2004). Traits associated with domestication arise, at least in part, from human selection and thereafter relate to ways that plants are cultivated and harvested. Gepts (2004) also indicated that a common feature among most of the domesticated plants is a marked genetic bottleneck.

Amazonia is the world's most important center of biodiversity (Clement, 1989). At the time of European conquest, hundreds of species were domesticated in Amazonia, including cassava, cacao, sweet potato, pineapple and numerous fruit trees and palms. The ample number of crops domesticated in this area, besides cultural diversity, make Amazonia a very interesting area to study the domestication of plants (Diamond, 2002; Clement *et al.*, 2010, 2016).

Annatto (*Bixa orellana* L.) is a crop domesticated in Amazonia (Clement *et al.*, 2016). Moreira *et al.* (2015) proposed that *B. urucurana* is the wild ancestor of cultivated annatto, *B. orellana*. They also accepted Pilger's proposal, published by Kuntz, that *urucurana* is a variety of *B. orellana* (*B. orellana* var. *urucurana* (Willd.) Kuntze ex Pilg.). Our observations in the field, combined with information obtained from farmers, indicate that wild and domesticated types cross and generate fertile offspring. According to the biological concept of

species, this is sufficient to determine that they are not two separate species. Therefore, from the time of Moreira *et al.* (2015) publication, we are now denominating wild annatto as *B. orellana* var. *urucurana*.

In this study, several collections were made in Amazonia and during the expeditions we identified both cultivated and wild annatto populations and observed a series of domestication syndromes. Cultivated annatto presents large fruits, a large quantity of seeds per fruit and a very strong coloring of the seeds, besides showing great variability in fruit colors and shapes (Figure 1.1). On the other hand, the wild type presents small fruits, with few seeds and with little coloration in the aryl of the seed, besides we observed only fruits of green color (Figure 1.2).

Annatto was domesticated by humans due to the dye that exists in the aril. This dye was widely used since ancient times. The Indians of tropical America made use of the colorful arils for body painting, dyeing of clothes, and especially for rituals (Plotkin, 1993). After the Europeans conquest, annatto became visible since Francisco de Orellana referred to annatto in one of his letters. The species was even named in honor of this navigator (Morton, 1960). Annatto was then widely used in the food and cosmetics industries in replace of synthetic pigments. The consumption of annatto increased since the prohibition of the use of synthetic dyes in some countries (Giuliano *et al.*, 2003; Nisar *et al.*, 2015). Brazil is the leader in annatto production, although Peru, Kenya, Colombia and other countries also grow annatto commercially. In Brazil, the states of Rondônia and São Paulo are the largest producers (Fabri, 2015).

Considering the economic and cultural importance of annatto, besides the fact that Brazil hosts great diversity for this crop, this work aimed to develop new microsatellite markers and generate information about its diversity, genetic structure, domestication and the reproductive system using microsatellite and cpDNA markers, plus phytochemical traits, studying both cultivated (*Bixa orellana*) and wild annatto (*B. orellana* var. *urucurana*), aiming to help to guide future research on breeding programs, as well as plan for the *in situ* and *ex situ* strategies for conservation of annatto in Brazil.

The specific objectives of this project are described below separately for each study.

Figure 1.1 Morphological variation on flowers, fruits and plant architecture of cultivated annatto (*Bixa orellana* L.)

Figure 1.2 Morphological variation on fruits and plant architecture of wild annatto (*Bixa orellana* var. *urucurana*)

1.1. Structure of the thesis and objectives

This thesis was organized in six studies presented in scientific manuscript format written in the English language. The studies were organized as follows:

"New microsatellite loci for annatto (*Bixa orellana*), a source of natural dyes from Brazilian Amazonia"

Objective: Develop a new set of microsatellite loci for *B. orellana* and its wild relative *B. orellana* var. *urucurana* as a tool to generate useful information for conservation strategies and population genetic studies.
This study was submitted and accepted for publication in the journal *Crop Breeding and Applied Biotechnology*.

"Highly structured genetic diversity of *Bixa orellana* var. *urucurana*, the wild ancestor of annatto, in Brazilian Amazonia"

Objective: Evaluate the genetic diversity and structure of wild populations using microsatellite markers and characterize the potential geographical range of this variety in northern South America using Ecoclimatic Niche Modeling (ENM).
This study was submitted to the journal *PlosOne*.

"Genetic diversity and structure in a major Brazilian annatto (*Bixa orellana*) germplasm bank revealed by microsatellites and phytochemical compounds"

Objective: Characterize the genetic diversity and structure of accessions from the annatto germplasm bank at the Agronomic Institute (IAC), Campinas, São Paulo, Brazil, using microsatellite markers and four phytochemical compounds.
This chapter was submitted and accepted for publication in the journal *Genetic Resources and Crop Evolution.*

"Genetic diversity and distribution of annatto (*Bixa orellana* L.) from Amazonia and Southeast of Brazil"

Objective: Characterize the amount and distribution of genetic diversity among cultivated accessions of cultivated annatto originated from the Amazonian region and also from Southern Brazil, in the State of São Paulo, using highly polymorphic microsatellite markers.

"Mixed mating system of annatto (*Bixa orellana*) determined with microsatellite markers"

Objective: Investigate the mating system of *Bixa orellana* in order to identify the levels of cross-fertilization and/or self-fertilization, and to better understand the genetic structure of progenies in a population of cultivated annatto from Brazilian Amazonia, providing information for conducting and establishing conservation and breeding programs.

"Domestication of annatto from *Bixa orellana* var. *urucurana*: Evidences based on SSR markers and chloroplast DNA"

Objective: In this study we aimed to answer the following questions: a) what are the levels of genetic diversity and genetic structure among cultivated annatto, compared to wild annatto? b) is there any evidence of gene flow, and therefore, of hybridization between wild and cultivated annatto? c) by the evidences gathered in this study, is it possible to suggest the most likely origin of domestication of annatto in Brazilian Amazonia?

REFERENCES

Badr, A. and El-Shazly, H. (2012) Molecular approaches to origin, ancestry and domestication history of crop plants: Barley and clover as examples. *J. Genet. Eng. Biotechnol.* **10**, 1–12.

Brown, A.H.D. (2010) Variation under domestication in plants: 1859 and today. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 2523–2530.

Clement, C.R. (1989) A center of crop genetic diversity in western Amazonia. *Bioscience* **39**, 624–631.

Clement, C.R., Cristo-Araújo, M. de, D'Eeckenbrugge, G.C., Pereira, A.A. and Picanço-Rodrigues, D. (2010) Origin and domestication of native Amazonian crops. *Diversity* **2**, 72–106.

Clement, C.R., Rodrigues, D.P., Alves-Pereira, A., Mühlen, G.S., Cristo-Araújo, M. De, Moreira, P.A., Lins, J. and Reis, V.M. (2016) Crop domestication in the upper Madeira River basin. *Bol. do Mus. Para. Emilio Goeldi de Ciencias Humanas* **11**, 193–205.

Diamond, J. (2002) Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707.

Fabri, E.G. (2015) Demanda por corantes naturais aquece mercado brasileiro de urucum. *Soc. Nac. Agric*.

Gepts, P. (2004) Crop domestication as a long term selection experiment, Oxford, UK: John Wiley & Sons, Inc.

Giuliano, G., Rosati, C. and Bramley, P.M. (2003) To dye or not to dye: biochemistry of annatto unveiled. *Trends Biotechnol.* **21**, 513–6.

Harlan, J. (1992) Crops and man, Madison: American Society of Agronomy.

Meyer, R.S. and Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852.

Moreira, P.A., Lins, J., Dequigiovanni, G., Veasey, E.A. and Clement, C.R. (2015) The domestication of Annatto (*Bixa orellana*) from *Bixa urucurana* in Amazonia. *Econ. Bot.* **69**, 127–135.

Morton, J.F. (1960) Can Annato (*Bixa orellana* L.) an old source of food color, meet new needs for safe dye. *Florida State Hortic. Soc.*, 301–309.

Nisar, N., Li, L., Lu, S., *et al.* (2015) Carotenoid metabolism in plants. *Mol. Plant* **8**, 68–82.

Plotkin, M. (1993) Tales of a Shaman's apprentice, New York: Penguin Books.

Terrell, J.E., Hart, J.P., Barut, S., *et al.* (2003) Domesticated landscapes: the subsistence ecology of plant and animal domestication. *J. Archaeol. Method Theory* **10**, 323–368.

2. NEW MICROSATELLITE LOCI FOR ANNATTO (*BIXA ORELLANA*), A SOURCE OF NATURAL DYES FROM BRAZILIAN AMAZONIA

Abstract

Annatto (*Bixa orellana* L.) is a tropical crop native to the Americas with Amazonia the likely center of origin of domestication. Annatto is important because it produces the dye bixin, which is widely used in the pharmaceutical, food, cosmetic and textile industries. A total of 32 microsatellite loci were isolated from a microsatellite-enriched genomic library, of which 12 polymorphic loci were used to characterize four populations of *B. orellana* and *B. orellana* var. *urucurana*, the wild relative. Higher genetic diversity estimates were detected for the wild populations when compared to the cultivated populations. Also, higher apparent outcrossing rates were found for the two wild populations when compared to the cultivated populations. These results indicate a mixed mating system for the species. All markers described herein have shown potential to be used in further studies evaluating the genetic diversity, population dynamics, domestication, improvement and conservation genetics of annatto.

Keywords: Bixaceae; Conservation Genetics; Genetic Diversity; Population Genetics; SSR

2.1 Introduction

Annatto (*Bixa orellana* L.; Bixaceae) is a tropical crop native to the Americas with Amazonia the likely center of origin of domestication (Arce 1999; Clement *et al.* 2010; Moreira *et al.* 2015). The wild ancestor of cultivated annatto has recently been identified as *B. orellana* var. *orellana* (Willd.) Kuntze ex Pilg. Hence (Moreira *et al.* 2015). Historical evidence indicates the extensive distribution and cultivation of annatto in the American tropics and subsequently its spread worldwide (Leal and Clavijo 2010). Brazil is the major producer of annatto and also hosts the greatest diversity of this species. Annatto is also produced by Peru, Kenya, Dominican Republic, Colombia, Jamaica, Costa Rica, Suriname and other countries in Asia (Akshatha *et al.* 2011). Annatto is commercially valuable due to its applications in the food and cosmetics industries, as a natural dye to be used instead of synthetic ones (Nisar *et al.* 2015). It is the second most important economic crop worldwide among all natural colorants, and has acquired notoriety for containing, in addition to dye, other important substances for human health, such as geranylgeraniol, tocotrienols and other carotenoids presenting antimicrobial, antioxidant and antiviral properties (Albuquerque and Meireles 2012), that can be used to treat human diseases, including leishmaniosis (Lopes *et al.* 2012). Annatto dye also has great importance in the Brazilian culture, since it is still used by indigenous tribes for body painting and dyeing of clothes (Plotkin 1993).

Microsatellites or simple sequence repeats (SSR) are important tools to assess genetic diversity and genetic structure of populations. They are widely present in eukaryotic genomes and very useful mainly because of their codominant inheritance, high polymorphism, high variability and suitability for automated allele sizing and cross-species transferability (Kalia *et al.* 2011; Vieira *et al.* 2016). Dequigiovanni *et al.* (2014) developed 10 polymorphic microsatellite markers for *B. orellana* L.; however, a large number of markers can increase the accuracy of population genetic parameter estimation. Thus, this study presents a new set of microsatellite loci for *B. orellana*, and its wild relative *B. orellana* var. *urucurana*, in order to generate useful information for conservation strategies and population genetics studies.

2.2 Material and methods

Ninety-eight individuals from four populations were analyzed in this study. Two populations are of cultivated annatto (*B. orellana*), one with 31 accessions located in São Francisco do Guaporé, Rondônia, Brazil (12°16'08.2"S; 63°25'27.1"W), and the other with 22 accessions located in Rondon do Pará, Pará, Brazil (4°44'43.4"S; 47°55'59.5"W). Markers developed in this study were also tested for cross amplification in two wild annatto (*B. orellana* var. *urucurana*) populations, one from Corumbiara, Rondônia, Brazil (13°39’15.80” S; 61°32’27.69”W), with 25 individuals, and another from Ariquemes, Rondônia, Brazil (9°55'30.5"S; 63°04'16.6"W), with 20 individuals.

Genomic DNA extraction from *Bixa orellana* and *B. orellana* var. *urucurana* samples was performed with the CTAB protocol (Doyle and Doyle 1990). A microsatellite-enriched library for *B. orellana* was developed following Billotte *et al.* (1999). Genomic DNA was digested with the enzime *Afa*I (Integrated DNA Technology-IDT, Coralville, USA) and the fragments resulting from digestion were linked to *Afa*21 and *Afa*25 adapters. Fragments were pre-amplified by Polymerase Chain Reaction (PCR) using the *Afa*21 adapter. Fragments containing repeats were selected with $(CTT)_{10}$, $(GT)_{10}$ and $(TA)_{10}$ biotinylated oligos, and recovered with streptavidin-coated magnetic particles (Sigma-Aldrich, St. Louis, USA). Enriched DNA fragments were amplified and cloned using the pGEM-T easy vector (Promega, Madison, USA) and transformed into XL1-BLUE *Escherichia coli* competent cells (Stratagene, Santa Clara, USA). Ninety-two positive clones were sequenced using universal T7 and SP6 primers with a BigDye v3.1 terminator kit on an ABI 3130*XL* Genetic Analyzer automated sequencer (Applied Biosystems, Foster City, USA). The selection of sequences

containing microsatellite repeats was performed using WebSat (Martins *et al.* 2009). We considered dinucleotides with more than six repeats, and trinucleotides, tetranucleotides and pentanucleotides with three or more repeats. Primer design was performed in PRIMER 3 (Rozen and Skaletsky 2000) considering sequences with 50-80% of GC content, with final products ranging from 130 to 350 base pairs (bp) and primers ranging in size from 18 to 22 bp. An M13 sequence tail was added to the 5' end of each forward primer following the Schuelke (2000) protocol.

Thirty-two microsatellite loci were characterized. PCRs were performed in a final volume of 10 μL, containing 20 ng of genomic DNA template, 1 U *Taq* DNA polymerase (Fermentas, Vilnius, Lithuania), 1X PCR buffer (10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.08% Nonidet P40), 0.25 mM each dNTP, 1.5 mM $MgCl_2$, 2.5 pmol of forward and M13 label primers (FAM, HEX or NED dyes) and 5 pmol of reverse primers. PCRs were carried out according to the Schuelke (2000) protocol, consisting of 94 °C (5 min), then 30 cycles at 94 °C (30 s)/Ta °C (45 s)/72 °C (45 s) [Ta= annealing temperature (Table 2.1)], followed by 8 cycles at 94 °C (30 s)/53 °C (45 s)/72 °C (45 s), and a final extension at 72 °C for 10 min. Quality of amplification was checked by electrophoresis in agarose gels (1.5%) stained with GelRed (Biotium, Hayward, USA). PCR products were visualized in an ABI 3130*XL* (Applied Biosystems, Foster City, USA) and allele sizes were scored using *GeneScan™–500 ROX® Size Standard* (Applied Biosystems, Foster City, USA) and analyzed with GENEMAPPER v4.0 software (Applied Biosystems, Foster City, USA). Descriptive statistics and Hardy–Weinberg Equilibrium (HWE) were calculated using *diveRsity* (Keenan *et al.* 2013) for R (R Core Team 2015). Genotypic disequilibrium between pairwise loci was estimated using *hierfstat* (Goudet 2005). Monte Carlo permutations of alleles between individuals and a Bonferroni correction (95%; $\alpha = 0.05$) were used to test if the estimates were significantly different from zero. The software micro-checker 2.2.1 (van Oosterhout *et al.* 2004) was used to identify possible genotyping errors resulting from stuttering or large allele dropout and the presence of null alleles within the microsatellite data set by performing 1000 randomizations.

Distribution of genetic variation within and among populations was evaluated using "locus-by-locus" AMOVA with GenAlEx version 6.5 (Peakall and Smouse 2012). Wright's F_{ST} was also used to estimate population differentiation and was calculated using GenAlEx. When populations are under Wright's equilibrium, the outcrossing rate is a function of the within-population inbreeding coefficient (Wright 1965). So, apparent outcrossing rate ($\hat{t}_a$)

was calculated for all populations according to Vencovsky (1994), with $\hat{t}_a$ = (1-*f*)/(1+*f*). Principal coordinate analysis (PCoA) was used to evaluate the dispersion of accessions with GenAlEx (Peakall and Smouse 2012).

Table 2.1 Description of 32 *Bixa orellana* microsatellite loci, including loci names, GenBank accession numbers, annealing temperatures (T_a), repeat motifs and size range of each locus

Locus	GenBank accession	T_a (°C)	Repeat motif	Size range (bp)
BorA3_2013	KJ922640	60.0	$(TC)_{16}$	216-252
BorA5_2013	KJ922646	62.0	$(TG)_{13}(GA)_{22}$	237-263
BorB1_2013	KJ922641	62.0	$(TG)_8N(TG)_{11}$	177-255
BorB5_2013	KJ922647	58.0	$(CA)_{19}$	259-291
BorC5_2013	KJ922648	60.0	$(CT)_{14}(CA)_9$	245-263
BorD1_2013	KJ922642	58.0	$(TG)_8(GA)_4$	167-177
BorD2_2013	KJ922643	62.0	$(TG)_{10}$	244-264
BorF9_2013	KJ922651	60.0	$(TA)_6(TG)_{10}$	290-310
BorG11_2013	KJ922655	56.0	$(TA)_6(CA)_{21}$	291-349
BorH3_2013	KJ922645	60.0	$(TG)_8$	236-244
BorH7_2013	KJ922652	56.0	$(AC)_8$	264-270
BorH10_2013	KJ922656	56.0	$(TA)_4(CA)_9$	295-327
BorC2_2013	KX017260	60.0	$(AG)_{30}$	209
BorG3_2013	KJ922644	58.0	$(GA)_{16}$	160
BorH1_2013	KX017261	62.0	$(AG)_{16}N(AG)_{10}$	205
BorC4_2013	KX017262	60.0	$(TTAT)_3$	151
BorE4_2013	KX017263	58.0	$(TC)_{10}(CA)_7$	194
BorF6_2013	KX017264	60.0	$(ATA)_4$	238
BorH6_2013	KJ922649	60.0	$(TC)_5(CA)_9$	206
BorB8_2013	KX017265	58.0	$(TG)_5$	204
BorC8_2013	KX017266	58.0	$(AG)_5$	202
BorD7_2013	KJ922650	58.0	$(GT)_9$	177
BorD9_2013	KX017267	58.0	$(AC)_8$	268
BorH9_2013	KX017268	60.0	$(AT)_5(GT)_7$	204
BorB11_2013	KX017269	60.0	$(AC)_{11}$	245
BorB12_2013	KX017270	60.0	$(AG)_{22}$	130
BorC10_2013	KJ922653	60.0	$(CA)_8$	222
BorC11_2013	KX017271	60.0	$(TG)_9(AG)_{17}$	272
BorE11_2013	KJ922654	60.0	$(CA)_8$	251
BorE12_2013	KX017272	60.0	$(AG)_{18}$	156
BorG10_2013	KX017273	58.0	$(GA)_{24}$	167
BorH11_2013	KX017274	58.0	$(CT)_{12}(AC)_9$	105

2.3 Results and discussion

Thirty-two loci amplified successfully (Table 2.1) from 92 positive clones sequenced from the library. Among these 32 loci, twelve were found to be polymorphic in *B. orellana*

and *B. orellana* var. *urucurana* populations (Table 2.2). This level of polymorphism (35%) was also observed in other studies with *Bixa orellana*. Dequigiovanni *et al.* (2014) found 10 polymorphic loci out of 25 loci evaluated. Micro-Checker detected no genotyping errors due to stuttering and large allele dropout. The analyses also showed that loci BorA5_2013, BorB1_2013, BorD1_2013, BorD2_2013, BorG11_2013 and BorH10_2013 might be affected by null-alleles in cultivated populations. This excess of homozygosity may be attributable to inbreeding. Therefore, none of the loci were excluded from the analyses.

Table 2.2. Genetic characterization of 12 polymorphic SSR loci in *Bixa orellana* (cultivated) and *B. orellana* var. *urucurana* (wild) populations. Genetic diversity described as number of alleles (*A*), observed (H_O) and expected (H_E) heterozygosities and inbreeding coefficient ($f= 1 -H_O/H_E$)

	Wild - Corumbiara				**Wild - Ariquemes**			
Locus	***A***	H_O	H_E	f	***A***	H_O	H_E	F
BorA3_2013	6	0.636	0.742	0.142	5	0.600	0.729	0.177
BorA5_2013	3	0.300	0.515	0.417	6	0.722	0.725	0.004
BorB1_2013	5	0.333	0.597	0.442	8	0.450	0.733	0.386
BorB5_2013	3	0.304	0.328	0.072	3	0.200	0.184	-0.088
BorC5_2013	7	0.792	0.792	0.000	5	0.900	0.646	-0.393
BorD1_2013	1	-	-	-	2	0.400	0.320	-0.250
BorD2_2013	8	0.792	0.814	0.028	4	1.000	0.591	-0.691
BorF9_2013	6	0.263	0.781	0.663	4	0.143	0.311	0.541
BorG11_2013	8	0.458	0.827	0.446	8	0.400	0.780	0.487
BorH3_2013	2	0.167	0.153	-0.091	4	0.200	0.597	0.661
BorH7_2013	2	0.600	0.471	-0.273	3	0.450	0.626	0.281
BorH10_2013	7	0.292	0.628	0.536	2	0.230	0.500	0.540
Mean	4.84	0.448	0.604	0.216	4.50	0.474	0.561	0.137
Total	58	-	-	-	54	-	-	-
	Cultivated - São Francisco do Guaporé				**Cultivated - Rondon do Pará**			
Locus	***A***	H_O	H_E	f	***A***	H_O	H_E	F
BorA3_2013	5	0.455	0.684	0.335	3	0.278	0.329	0.155
BorA5_2013	3	0.194	0.629	0.692	2	0.143	0.278	0.486
BorB1_2013	6	0.226	0.713	0.683	3	0.400	0.629	0.364
BorB5_2013	4	0.133	0.336	0.603	4	0.316	0.582	0.457
BorC5_2013	5	0.308	0.553	0.444	2	0.381	0.444	0.143
BorD1_2013	1	-	-	-	2	0.000	0.165	1.000
BorD2_2013	2	0.000	0.391	1.000	1	-	-	-
BorF9_2013	4	0.182	0.498	0.635	5	0.438	0.678	0.354
BorG11_2013	5	0.120	0.730	0.836	2	0.000	0.100	1.000
BorH3_2013	1	-	-	-	1	-	-	-
BorH7_2013	3	0.133	0.472	0.717	1	-	-	-
BorH10_2013	5	0.133	0.526	0.746	3	0.381	0.571	0.333
Mean	3.67	0.188	0.553	0.669	2.41	0.259	0.419	0.476
Total	44	-	-	-	29	-	-	-

Polymorphic loci were used to calculate descriptive statistics for each population (Table 2.2). For the wild *B. orellana* var. *urucurana* the number of alleles per locus varied from 1 to 8. A lower number of alleles per loci was found for cultivated annatto, varying from 1 to 6 (Table 2.2). The average observed (H_O) and expected heterozygosities (H_E) were also higher in the wild than in the cultivated populations, with higher values observed for H_E than H_O in both wild and cultivated populations. As a result of this, local inbreeding coefficients were high in all populations (Table 2.2). Similar results for cultivated accessions were observed by Dequigiovanni *et al.* (2014) (A = 3.8; H_O = 0.54; H_E= 0.63).

Higher levels of genetic diversity in wild compared to cultivated crops has also been found in other crops, due to bottleneck effects during domestication, such as tepary beans (*Phaseolus acutifolius*) (Blair *et al.* 2012; Gujaria-Verma *et al.* 2016), common beans *(P. vulgaris)* (Bitocchi *et al.* 2013), apricot (*Prunus armeniaca)* (Bourguiba *et al.* 2012), sunflower *(Helianthus annuus)* (Mandel *et al.* 2011). However, this is not always the case, since in some crops a decrease of genetic diversity during domestication did not occur, as in carrot (*Daucus carota* subsp. sativus) (Iorizzo *et al.* 2013).

Deviation from Hardy–Weinberg equilibrium (HWE) was tested for all loci and populations. Ten loci were found deviating from HWE due to excess heterozygosity for *B. orellana* and five loci for *B. orellana* var. *urucurana*. Deviations from HWE may occur because *B. orellana* presents a mixed mating system and can tolerate both autogamy and allogamy (Rivera-Madrid *et al.* 2006; Valdez-Ojeda *et al.* 2010; Joseph *et al.* 2012). Similarly, Dequigiovanni *et al.* (2014) found deviations of HWE in eight out of ten loci analyzed. No significant linkage disequilibrium was detected for each pair of loci tested after Bonferroni correction.

The apparent outcrossing rates estimated for all populations in this study indicated a mixed mating system for annatto, with much higher outcrossing rates observed for the two wild populations ($\hat{t}_a$ = 0.644 for Corumbiara; $\hat{t}_a$ = 0.759 for Ariquemes) than for cultivated annatto ($\hat{t}_a$ = 0.198 for São Francisco do Guaporé/RO; $\hat{t}_a$ = 0.355 for Rondón do Pará, PA). Also, it is interesting to mention that the farmer from São Francisco do Guaporé had a much more uniform commercial annatto plantation than the Rondón do Pará farmer, a more traditional type of farmer, which is reflecting in a lower outcrossing rate for the first one.

The AMOVA analysis identified higher proportion of genetic variation within populations (68%) than among populations (29%, F_{ST} = 0.317, P < 0.001), which is still quite high and suggest that subdivision has a great impact on the genetic diversity. However, only

2% of total variation was attributable to differences between wild and cultivated populations, showing there must be considerable gene flow between these two types of populations, especially in Rondônia (Figure 2.1). *F*-statistics (F_{IS} = 0.366; F_{ST} = 0.367; F_{IT} = 0.597) also confirmed high levels of genetic structure. The cultivated population in Rondon do Pará is the most divergent, apparently indicating isolation by distance, while the two wild and one cultivated populations in Rondônia show some to considerable gene flow.

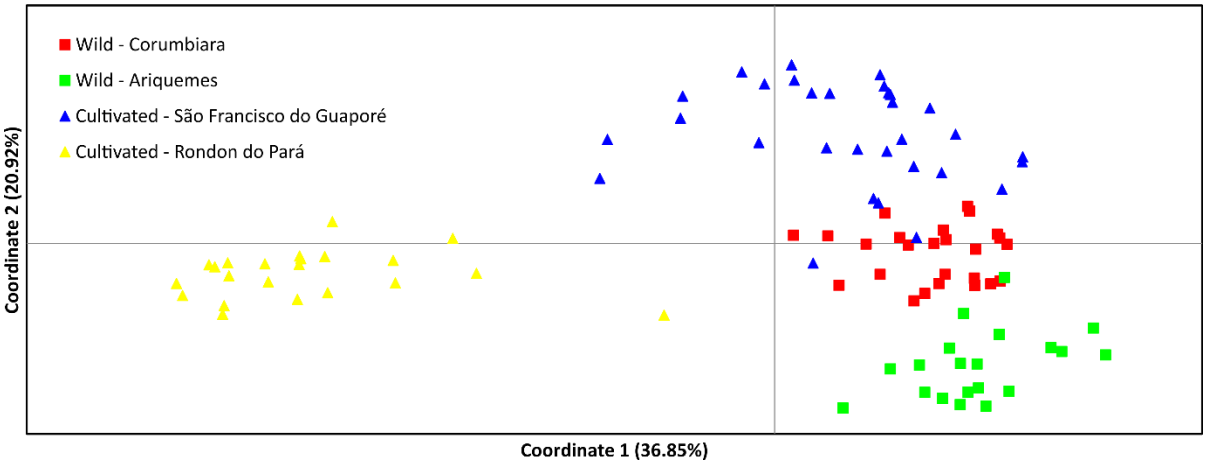

Figure 2.1 Principal coordinate analysis of the dispersion of two cultivated populations of *Bixa orellana* (Rondon do Pará and São Francisco do Guaporé) and two wild populations of *B. orellana* var. *urucurana* (Corumbiara and Ariquemes), using 12 newly developed microsatellite markers.

In conclusion, the 12 polymorphic loci reported in this study have proven to be powerful tools for assessing genetic diversity, genetic structure, and also for domestication studies in *B. orellana* and *B. orellana* var. *urucurana*. Higher levels of genetic diversity and outcrossing rates were found for the wild populations when compared to the cultivated populations. Also, most of the variation found for SSR markers is located within populations, which apparently have a mixed mating system. Loci that presented monomorphism in these populations may present polymorphism in other populations and, therefore, should not be discarded.

REFERENCES

Akshatha, V., Giridhar, P. and Ravishankar, G.A. (2011) Morphological diversity in *Bixa orellana* L. and variations in annatto pigment yield. *J. Hortic. Sci. Biotechnol.* **86**, 319–324.

Albuquerque, C.L.C. and Meireles, M.A.A. (2012) Defatting of annatto seeds using supercritical carbon dioxide as a pretreatment for the production of bixin: Experimental, modeling and economic evaluation of the process. *J. Supercrit. Fluids* **66**, 86–94.

Arce, J. (1999) El achiote *Bixa orellana* L. cultivo promisorio para el trópico. In Earth., p. 149.

Billotte, N., Lagoda, P., Risterucci, A. and Baurens, F. (1999) Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. *Fruits* **54**, 277–288.

Bitocchi, E., Bellucci, E., Giardini, A., et al. (2013) Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytol.* **197**, 300–313.

Bourguiba, H., Audergon, J.-M., Krichen, L., et al. (2012) Loss of genetic diversity as a signature of apricot domestication and diffusion into the Mediterranean Basin. *BMC Plant Biol.* **12**, 49.

Clement, C.R., Cristo-Araújo, M. de, D'Eeckenbrugge, G.C., Pereira, A.A. and Picanço-Rodrigues, D. (2010) Origin and domestication of native Amazonian crops. *Diversity* **2**, 72–106.

Dequigiovanni, G., Ramos, S.L.F., Zucchi, M.I., Bajay, M.M., Pinheiro, J.B., Fabri, E.G., Bressan, E.A. and Veasey, E.A. (2014) Isolation and characterization of microsatellite loci for *Bixa orellana*, an important source of natural dyes. *Genet. Mol. Res.* **13**.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of Plant DNA from fresh tissue. *Focus (Madison).* **12**, 13–15.

Galeano, C.H., Cortés, A.J., Fernández, A.C., Soler, Á., Franco-Herrera, N., Makunde, G., Vanderleyden, J. and Blair, M.W. (2012) Gene-Based Single Nucleotide Polymorphism Markers for Genetic and Association Mapping in Common Bean. *BMC Genet.* **13**, 48.

Goudet, J. (2005) hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186.

Gujaria-Verma, N., Ramsay, L., Sharpe, A.G., Sanderson, L.-A., Debouck, D.G., Tar 'an, B., Bett, K.E., Tar'an, B. and Bett, K.E. (2016) Gene-based SNP discovery in tepary bean (Phaseolus acutifolius) and common bean (P. vulgaris) for diversity analysis and comparative mapping. *BMC Genomics* **17**, 239.

Iorizzo, M., Senalik, D.A., Ellison, S.L., et al. (2013) Genetic structure and domestication of carrot (Daucus carota subsp. sativus) (Apiaceae) 1. *Am. J. Bot.* **100**, 930–938.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R. and Dhawan, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177**, 309–334.

Keenan, K., McGinnity, P., Cross, T.F., Crozier, W.W. and Prodöhl, P.A. (2013) diveRsity : An R package for the estimation and exploration of population genetics parameters and their associated errors O'Hara, R.B., ed. *Methods Ecol. Evol.* **4**, 782–788.

Leal, F. and Clavijo, C.M. (2010) Acerca de la história, taxonomia, botánica y usos de *Bixa orellana* L. *Rev. Unell. Cienc. Tec* **1**, 78–86.

Lopes, M.V., Desoti, V.C., Caleare, A.D.O., Ueda-Nakamura, T., Silva, S.O. and Nakamura, C.V. (2012) Mitochondria superoxide anion production contributes to geranylgeraniol-induced death in *Leishmania amazonensis*. *Evidence-based Complement. Altern. Med.* **2012**, 298320.

Mandel, J.R., Dechaine, J.M., Marek, L.F. and Burke, J.M. (2011) Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, Helianthus annuus L. *Theor. Appl. Genet.* **123**, 693–704.

Martins, W.S., Lucas, D.C.S., Neves, K.F. de S. and Bertioli, D.J. (2009) WebSat--a web software for microsatellite marker development. *Bioinformation* **3**, 282–3.

Monzote, L., Montalvo, A.M., Almanonni, S., Scull, R., Miranda, M. and Abreu, J. (2006) Activity of the essential oil from *Chenopodium ambrosioides* grown in Cuba against *Leishmania amazonensis*. *Chemotherapy* **52**, 130–6.

Moreira, P.A., Lins, J., Dequigiovanni, G., Veasey, E.A. and Clement, C.R. (2015) The domestication of Annatto (*Bixa orellana*) from *Bixa urucurana* in Amazonia. *Econ. Bot.* **69**, 127–135.

Nisar, N., Li, L., Lu, S., *et al.* (2015) Carotenoid metabolism in plants. *Mol. Plant* **8**, 68–82.

Oosterhout, C. Van, Hutchinson, W.F., Wills, D.P.M. and Shipley, P. (2004) MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4**, 535–538.

Peakall, R. and Smouse, P.E. (2006) GENALEX 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* **6**, 288–295.

Plotkin, M. (1993) Tales of a Shaman's apprentice, New York: Penguin Books.

R Core Team (2015) R: A language and environment for statistical computing.

Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–86.

Schuelke, M. (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18**, 233–234.

Valdez-Ojeda, R., Quiros, C.F., Lourdes Aguilar-Espinosa, M. de, Rivera-Madrid, R., Aguilar-Espinosa, M. de L. and Rivera-Madrid, R. (2010) Outcrossing rates in annatto determined by sequence-related amplified polymorphism. *Agron. J.* **102**, 1340–1345.

Vencovsky, R. (1994) Variance of an estmatve of the outcrossing rate. *Rev. Bras. Genética* **17**, 349–351.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. (2016) Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **39**, 312–328.

Wright, S. (1965) The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution (N. Y).* **19**, 395.

3. HIGHLY STRUCTURED GENETIC DIVERSITY OF *BIXA ORELLANA* VAR. *URUCURANA*, THE WILD ANCESTOR OF ANNATTO, IN BRAZILIAN AMAZONIA

Abstract

Annatto (*Bixa orellana* L.) is a tropical American crop, commercially valuable due to its application in the food and cosmetics industries as a natural dye. The wild ancestor of cultivated annatto is *B. orellana* var. *urucurana*. Although never cultivated, this variety occurs in open forests and anthropogenic landscapes, and is always associated with riparian environments. In this study, we evaluated the genetic diversity and structure of *B. orellana* var. *urucurana* populations in Brazilian Amazonia using 16 microsatellite loci. We used Ecological Niche Modeling (ENM) to characterize the potential geographical range of this variety in northern South America. We analyzed 170 samples from 10 municipalities in the states of Rondônia, Pará and Roraima. A total of 194 alleles was observed, with an average of 12.1 alleles per locus. Higher levels of expected (H_E) than observed (H_O) heterozygosities were found for all populations. Bayesian analysis, Neighbor-Joining dendrograms and PCAs suggest the existence of three strongly structured groups of populations. A strong and positive correlation between genetic and geographic distances was found, suggesting that genetic differentiation might be caused by geographic isolation. From species distribution modelling, we detected that South Rondônia, Madre di Dios River basin, Llanos de Mojos, Llanos de Orinoco and eastern Ecuador are highly suitable areas for wild annatto to occur, providing additional targets for future exploration and conservation. Climatic adaptation analyses revealed strong differentiation among populations, suggesting that precipitation plays a key role in wild annatto's current and potential distribution patterns.

Keywords: SSR; Ecological Niche Modeling; Amazonia; Population Genetics; Annatto

3.1. Introduction

Annatto (*Bixa orellana* L.) is a tropical American crop (Arce, 1999), which probably originated in Amazonia (Schultes, 1984; Sandy-Cuen and Becerra, 2003; Clement *et al.*, 2010). Annatto is commercially valuable due to its application in the food and cosmetics industries, as a natural dye to be used instead of synthetic ones (Nisar *et al.*, 2015). Five species are recognized in the genus *Bixa* (*Bixa orellana* L., *B. arborea* Huber, *B. excelsa* Gleason & Krukoff, *B. platycarpa* Ruiz & Pav. ex G.Don, and *B. urucurana* Willd.) (Baer, 1976), which belongs to the Bixaceae family. The only cultivated species of the genus, *B. orellana*, is an evergreen shrub that is confined to the frost-free tropics (Schultes, 1984; Akshatha *et al.*, 2011). An important distinction among the five species is growth habit, which can be either a tree or a shrub. *B. orellana* and *B. urucurana* are shrubs, while *B. arborea, B. excels*a and *B. platycarpa* are trees (Baer, 1976; Moreira *et al.*, 2015). Ducke (1946)

hypothesized that *B. excelsa* might have been the wild ancestor of *B. orellana*, which was accepted by Schultes (1984) and Meyer *et al* (2012). However, *B. excelsa* is a tree and it is unlikely that domestication during the Holocene would transform all known populations into a shrub (Moreira *et al.*, 2015). Analysis of the domestication syndrome in the shrubby *Bixas* allowed Moreira *et al* (2015) to propose that *B. urucurana* is the wild ancestor of cultivated annatto, *B. orellana*. They also accepted Pilger's proposal, published by Kuntz (1925), that *urucurana* is a variety of *B. orellana* (*B. orellana* var. *urucurana* (Willd.) Kuntze ex Pilg.). The word 'urucurana' is derived from the Tupi language in which "rana" means false, and is often attributed to wild populations of a species with domesticated populations (Moreira *et al.*, 2015).

B. orellana var. *urucurana* occurs in open forests and anthropogenic landscapes, although never cultivated and is always associated with riparian environments (Baer, 1976; Moreira *et al.*, 2015). While cultivated annatto always produces abundant pigment around its seeds, urucurana contains variable amounts. In areas where they co-exist, gene flow between them results in changes in pigment production, especially in the domesticated types (Moreira *et al.*, 2015). The exact location where annatto was first domesticated is still unclear, mainly because of the wide distribution of variety *urucurana* in northern South America (Moreira *et al.*, 2015).

In order to make reasoned decisions about sampling procedures to preserve high levels of genetic diversity, researchers must know how genetic variation is organized and distributed throughout the geographic range of a species (Loveless and Hamrick, 1984). The assessment of genetic diversity and structure within and among populations of plants is generally performed using molecular markers. Microsatellites or SSRs (Simple Sequence Repeats) are among the most important molecular markers because they are abundant, co-dominant, with ample distribution in the genome, and generally neutral and highly polymorphic (Vieira *et al.*, 2016). Hence, SSR markers are important tools to assess genetic diversity and genetic structure of populations, especially for wild species (Vieira *et al.*, 2016). There are very few genetic studies in annatto (Valdez-Ojeda *et al.*, 2008; Valdez-Ojeda *et al.*, 2010; Dequigiovanni *et al.*, 2014) and no studies have yet evaluated population structure and genetic diversity of wild populations of annatto (*B. orellana* var. *urucurana*).

In this study, 170 samples from 10 populations of wild annatto in Brazilian Amazonia were collected and analyzed using 16 SSR markers, in order to answer the following questions: a) what are the levels of genetic diversity in these populations; b) what are the genetic relationships among these populations; and c) is genetic diversity geographically

structured across these populations? We used two approaches to answer these questions, combining Ecological Niche Modeling (ENM) and neutral genetic markers. ENM methods approximate an envelope for the environmental requirements of a taxon from a set of its occurrence localities, summarizing environmental variation across those landscapes to develop a quantitative picture of the potential distribution of the species. They have provided a powerful tool for investigating the ecology and distribution of both plant and animal species (Coppens d'Eeckenbrugge and Lacape, 2014; Peterson, 2003). Therefore, ENM was utilized to characterize the potential geographical range of *B. orellana* var. *urucurana* in northern South America, based on these Brazilian Amazonian populations.

3.2. Material and methods

3.2.1. Plant material

During our field work from 2009 to 2015, 170 samples of wild annatto (*Bixa orellana* var. *urucurana*) were collected in 10 municipalities in the states of Rondônia, Pará and Roraima, in Brazilian Amazonia (Table 3.1; Figure 3.1). From each plant, leaf samples were collected and stored in plastic bags containing silica gel. Some of the samples were stored in CTAB gels (3% w/v Cetyl Trimethyl Ammonium Bromide and 35% w/v NaCl). Each collection site was registered using Global Positioning System (GPS).

Table 3.1 Geographic location of the 10 populations of *Bixa orellana var. urucurana* collected in Brazilian Amazonia and used in this study, including sampling size (N), latitude and longitude (in decimal degrees).

Population ID / Municipality, State*	**N**	**Latitude**	**Longitude**
1 – Cabixi, RO	45	-13.48838	-60.60608
2 – Cerejeiras, RO	32	-13.17171	-60.80942
3 – Corumbiara, RO	26	-12.99158	-60.92277
4 – São Francisco do Guaporé, RO	9	-11.72616	-62.34804
5 – Jí-Paraná, RO	19	-11.49186	-62.41528
6 – Ariquemes, RO	18	-09.92515	-63.07129
7 – Mucajaí, RR	5	2.37	-61.44
8 – Monte Alegre, PA	5	-1.981198	-54.16811
9 – Almeirin, PA	4	-1.241724	-53.04789
10 – Bom Jesus do Tocantins, PA	7	-5.103889	-48.548889
Total	170	-	-

* States are: PA – Pará, RO – Rondônia, RR – Roraima

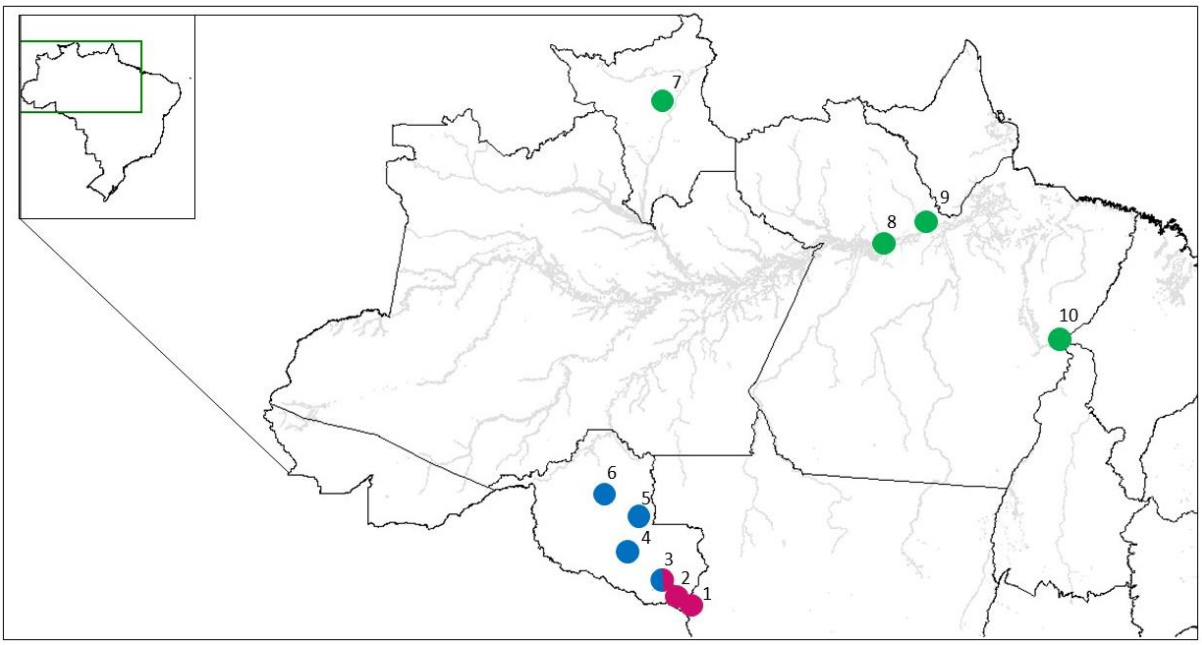

Figure 3.1 Geographic distribution of collection localities of 10 *Bixa orellana* var. *urucurana* populations and their assignments to the most likely number of clusters (K = 3) identified using *Structure*. Numbers represent collected populations: 1 – Cabixi, RO; 2 – Cerejeiras, RO; 3 – Corumbiara, RO; 4 – São Francisco do Guaporé, RO; 5 – Jí-Paraná, RO; 6 – Ariquemes, RO; 7 – Mucajaí, RR; 8 – Monte Alegre, PA; 9 – Almeirin, PA; 10 – Bom Jesus do Tocantins, PA.

3.2.2. DNA isolation, PCR amplification and genotyping of SSRs

Total genomic DNA was extracted from young leaves following Doyle and Doyle (1990) with CTAB 3%. DNA concentration was determined by comparison with known concentrations of standard DNA (lambda DNA, Invitrogen) during electrophoresis in agarose gels (1%) stained with GelRed (Biotium) under ultraviolet light.

Sixteen SSR markers developed for *B. orellana* (Dequigiovanni *et al.*, 2014) (Chapter 2, Dequigiovanni *et al.*, submitted) were 20 ng of DNA template, 1X polymerase chain reaction buffer (Fermentas, Vilnius, Lithuania), 0.25 mM of each dNTP, 1.5 mM of $MgCl_2$, 2.5 pmol of forward and M13 labeled primers (FAM, HEX or NED dyes), 5 pmol of reverse primers and 1 U of *Taq* DNA polymerase (Fermentas).

PCR was carried out according to Schuelke (2000) in a two-step process as follows: the first step consisted of an initial denaturing step of 94 °C for 5 min, followed by 30 cycles of 94 °C for 30 s, annealing temperature for 45 s, and 72 °C for 45 s. The second step consisted of 8 cycles at 94 °C for 30 s, 53 °C for 45 s and 72 °C for 45 s, and a final extension at 72 °C for 10 min. Quality of PCR products was checked by electrophoresis in agarose gels (1.5%) stained with GelRed (Biotium) under ultraviolet light. Capillary electrophoresis involved multiplexed marker panels, based on expected allele size, with two to three markers

with at least 80 bp size differences. Fragment separation and detection were performed on an ABI Prism 3130xl capillary sequencer (Applied Biosystems) with the aid of GeneScan 500 Rox Size Standard (Applied Biosystems). DNA fragment sizes were determined using GeneMapper software (Applied Biosystems).

3.2.3. SSR data analysis

Possible clusters of wild annatto were hypothesized using a Bayesian analysis with *Structure* software (Pritchard *et al.*, 2000). The number of clusters (*K*) was estimated by performing ten independent runs for each *K* (from 1 to 10, the number of geographic locations, hereafter called populations for convenience), using 1,000,000 MCMC repetitions and a 200,000 burn-in period. Correlated allele frequencies and admixture were assumed. The most likely number of clusters was evaluated with the ad hoc method of Evanno *et al* (2005).

Based on the original populations and the clusters identified by *Structure*, we estimated parameters for genetic diversity, including number of alleles per locus (*A*), effective allele number (Ne), allelic richness (A_R) (El Mousadik and Petit, 1996), observed (H_O) and expected heterozygosity (H_E), in addition to Wright's (1965) inbreeding coefficient (*f*). The apparent outcrossing rate ($\hat{t}_a$) was estimated considering the inbreeding coefficient (*f*) for each population (Vencovski, 1994), so that $\hat{t}_a$ = (1-*f*)/(1+*f*). The genetic diversity indices *A*, Ne, H_O and H_E were estimated with *GenAlEx 6.5* (Peakall and Smouse, 2012), and estimations of A_R and *f*, with confidence intervals based upon 1,000 bootstrap replicates, were obtained with *diveRsity* (Keenan *et al.*, 2013) and *poppr* (Kamvar *et al.*, 2014) for R (R Core Team, 2015).

In order to represent the relationships between individuals and populations, neighbor-joining (Saitou and Nei, 1987) dendrograms were constructed with *Phylip 3.5* (Felsenstein, 2005), based on Cavalli-Sforza and Edwards' (Cavalli-Sforza and Edwards, 1967) Chord distance (D_{CE}) obtained with *MSA 4.05* (Dieringer and Schlötterer, 2003). The Chord distance is a geometric distance and performs well for the reconstruction of relationships among populations (Reif *et al.*, 2005). Confidence of relationships was assessed with 1,000 bootstrap replicates. Final trees were formatted in *FigTree 1.4.1* (http://tree.bio.ed.ac.uk/software/figtree/). A principal coordinate analysis was used to visualize the dispersion of samples as a function of genetic variation using *GenAlEx 6.5* (Peakall and Smouse, 2012).

Hierarchical distribution of genetic variation within and among populations of wild annatto, and within and among groups according to the *Structure* analysis was evaluated using "locus-by-locus" AMOVA with *GenAlEx 6.5* (Peakall and Smouse, 2012). Gene flow (N_m) among populations was estimated by calculating $N_m = (1 - F_{ST})/4F_{ST}$ (Slatkin and Barton, 1989). In addition, the Mantel test was used to evaluate the correlation between Nei's genetic distance and geographic distance (km) among populations using *Adegenet* (Jombart and Ahmed, 2011) for R (R Core Team, 2015). Significance was assessed by conducting 9999 permutations.

3.2.4. Potential distribution of B. orellana var. urucurana

The potential distribution of *B. orellana* var. *urucurana* was estimated using the maximum entropy algorithm of *Maxent* v. 3.3.3e (Phillips *et al.*, 2006). Maxent estimates the potential distribution of a taxa from a maximum entropy probability distribution using presence-only data (Elith *et al.*, 2006). The resulting model is a geographical projection of habitat suitability for the target species where values close to 0 indicate sites that do not match with the niche requirements of the species, and values close to 1 indicate sites that fully match the niche requirements. A total of 184 presence-only records were compiled from field work and from georeferenced herbarium data extracted from the speciesLink project (http://splink.cria.org.br) and Global Biodiversity Information Facility (GBIF) portal. All geographic coordinates were manually verified and incomplete or imprecise records were discarded. For each occurrence record, we obtained 19 bioclimatic variables derived from monthly temperature and rainfall from the *WORLDCLIM* database with resolutions of 2.5' (Hijmans *et al.*, 2005). Fifteen model replicates were run with 75% of occurrences used for calibration and different subsets (25%) used for validation. A logistic threshold value of 10 percentile training presence was retained to separate climatically favorable areas from marginally fit areas. The accuracy of model prediction was evaluated using the area under the curve (AUC), where 1 was the maximum prediction and 0.5 suggested a random prediction (Fielding and Bell, 1997). Permutation procedure was used to define contributions of the variables to the models.

Because we observed a high correlation between genetic and geographic distances, we tested the assumption that most of the variability is due to environmental factors. To compare

the environmental characteristics of the different areas, we performed principal components analysis (PCA) with *ade4* (Dray and Dufour, 2007) for R (R Core Team, 2015).

The 19 bioclimatic variables are: BIO1 = Annual Mean Temperature; BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp)); BIO3 = Isothermality (BIO2/BIO7) (* 100); BIO4 = Temperature Seasonality (standard deviation *100); BIO5 = Max Temperature of Warmest Month; BIO6 = Min Temperature of Coldest Month; BIO7 = Temperature Annual Range (BIO5-BIO6); BIO8 = Mean Temperature of Wettest Quarter; BIO9 = Mean Temperature of Driest Quarter; BIO10 = Mean Temperature of Warmest Quarter; BIO11 = Mean Temperature of Coldest Quarter; BIO12 = Annual Precipitation; BIO13 = Precipitation of Wettest Month; BIO14 = Precipitation of Driest Month; BIO15 = Precipitation Seasonality (Coefficient of Variation); BIO16 = Precipitation of Wettest Quarter; BIO17 = Precipitation of Driest Quarter; BIO18 = Precipitation of Warmest Quarter; BIO19 = Precipitation of Coldest Quarter.

3.3. Results

3.3.1. Genetic diversity

All the 16 SSR markers were polymorphic, with a total of 194 alleles. The number of alleles per locus ranged from four (BorH3) to 19 (BorG11) with an average of 12 alleles per locus (Table 3.2). The observed heterozygosity (H_O) values for each locus ranged from 0.158 to 0.712 across loci, with a mean of 0.385, while the expected heterozygosity (H_E) ranged from 0.292 to 0.679, with a mean of 0.520. All loci had heterozygote deficits greater than 10%. The mean Shannon diversity index (I) was 0.975, ranging from 0.472 to 1.354 (Table 3.2).

Table 3.2 Genetic diversity estimates for 16 microsatellite (SSR) loci used to analyze wild annatto (*Bixa orellana* var. *urucurana*) collected in Brazilian Amazonia. Genetic diversity is described as number of alleles (*A*), observed (H_O) and expected (H_E) heterozygosities, and Shannon's information index (*I*).

Loci	***A***	$\boldsymbol{H_O}$	$\boldsymbol{H_E}$	***I***
BorA2	5	0.354	0.438	0.712
BorA3	17	0.391	0.463	0.916
BorA5	12	0.401	0.525	0.968
BorB1	17	0.311	0.641	1.268
BorB4	17	0.352	0.609	1.111
BorB5	12	0.379	0.458	0.834
BorB12	14	0.421	0.508	1.015
BorC5	12	0.668	0.712	1.286
BorD1	8	0.357	0.423	0.752
BorD2	10	0.468	0.434	0.841
BorF9	12	0.231	0.629	1.220
BorG4	14	0.465	0.569	1.102
BorG11	19	0.426	0.679	1.354
BorH3	4	0.158	0.292	0.472
BorH7	5	0.494	0.443	0.692
BorH10	16	0.239	0.540	1.066
Total	194	-	-	-
Mean	12.125	0.382	0.522	0.975

Genetic diversity estimates of the 10 populations showed a mean number of alleles per locus ($\bar{A}$) of 3.86 (Table 3.3), while allelic richness (A_R) had a mean value of 1.54 and the mean number of effective alleles per locus (N_E) was 2.64. Forty-one private alleles were observed, representing 21% of all alleles. The population from Bom Jesus do Tocantins (Population 10 in Table 3.1 and Figure 3.1) showed the highest number of private alleles (15) (Table 3.3). The mean values of observed (H_O) and expected (H_E) heterozygosities for all populations were 0.382 and 0.522, respectively. Significant inbreeding coefficients (*f*) were detected in most populations, ranging from 0.047 to 0.565. The mean apparent outcrossing rate ($\hat{t}_a$) was 0.609. When disregarding the populations with small sampling sizes, such as Monte Alegre (N = 5), Mucajaí (N = 5) and Almeirin (N = 4), the mean value of this parameter increased to 0.690.

Table 3.3 Genetic parameters estimated for 10 populations of *Bixa orellana* var. *urucurana* and for the three groups identified by the Structure analysis, including mean number of alleles per locus *($\bar{A}$)*, alellic richness *(A_R)*, mean number of effective alleles per locus *(N_E)*, observed *(H_O)* and expected *(H_E)* heterozygosity, local inbreeding coefficient *($f = 1 - H_O/H_E$)*, and apparent outcrossing rate ($\hat{t}_a$).

Population	$\bar{A}$ (private aleles)	A_R	N_E	H_O	H_E	f	$\hat{t}_a$
1. Cabixi, RO	5.813 (1)	1.582	3.000	0.449	0.577	0.217*	0.643
2. Cerejeiras, RO	4.875 (1)	1.558	2.845	0.442	0.554	0.181*	0.693
3. Corumbiara, RO	4.750 (1)	1.577	3.127	0.445	0.571	0.196*	0.672
4. S. F. do Guaporé, RO	3.625 (3)	1.564	2.512	0.437	0.548	0.168*	0.712
5. Jí-Paraná, RO	5.000 (3)	1.645	3.379	0.426	0.636	0.309*	0.528
6. Ariquemes, RO	3.875 (3)	1.565	2.566	0.453	0.553	0.166*	0.715
7. Mucajaí, RR	3.438 (9)	1.631	2.816	0.268	0.596	0.565*	0.278
8. Monte Alegre, PA	1.813 (2)	1.338	1.653	0.313	0.321	0.047	0.910
9. Almeirin, PA	2.063 (3)	1.353	1.865	0.146	0.327	0.508*	0.326
10. B.J.Tocantins, PA	3.313 (15)	1.535	2.596	0.471	0.516	0.081*	0.850
Mean	3.856	1.535	2.636	0.385	0.519	0.243	0.609
Structure Group							
South RO**	6.562	4.765	3.464	0.466	0.605	0.207*	0.657
Central RO	6.876	5.478	4.159	0.441	0.706	0.367*	0.463
PA and RR	6.125	5.558	4.221	0.337	0.701	0.528*	0.309
Mean	6.521	5.270	3.948	0.414	0.670	0.367	0.463

* significant based upon 1,000 bootstrap replicates
** State abbreviations are PA – Pará, RO – Rondônia, RR – Roraima

3.3.2. Genetic structure

The 170 wild annatto samples of 10 populations were grouped into genetic clusters by the *Structure* simulations, with a clear ΔK maximum at $K = 3$, and possible structure at $K = 2$ and $K = 7$ (Figures 3.2 and 3.4). According to $K = 3$, group I (hereafter South RO) included the populations from Cabixi, Cerejeiras and Corumbiara in the Guaporé River basin in southern Rondônia State. Group II (hereafter Central RO) included the populations from Ariquemes and Jí-Parana in the Jí-Parana River basin, and São Francisco do Guaporé, from the Guaporé River basin, located in central Rondônia State. The groups of South RO and Central RO meet and mix at Corumbiara, in southern Rondônia. Group III (hereafter PA and RR) included all the other populations, both north of the Amazon River in Roraima and Pará, and south of the Amazon River in eastern Pará (Figure 3.1). In $K = 2$, the South RO and Central RO groups were clustered together. At $K = 7$, the PA and RR group was subdivided, with the north of the Amazon River in one group, and eastern Pará in another group; groups South RO and Central RO were also subdivided, confirming the high diversity observed in both groups (Table 3.3).

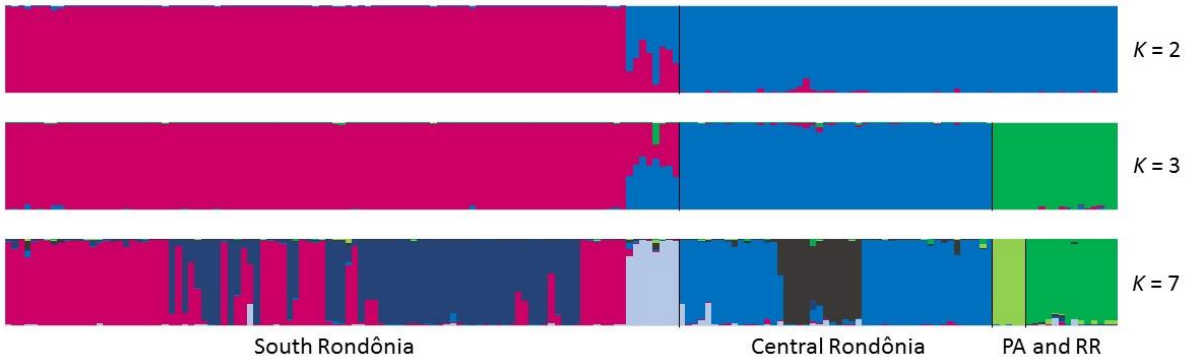

Fig 3.2 – Assignment of each of 170 *B. orellana* var. *urucurana* plants collected in Brazilian Amazonia to groups simulated by *Structure* at *K* = 2, *K*= 3 and *K* = 7 based on 16 SSR loci.

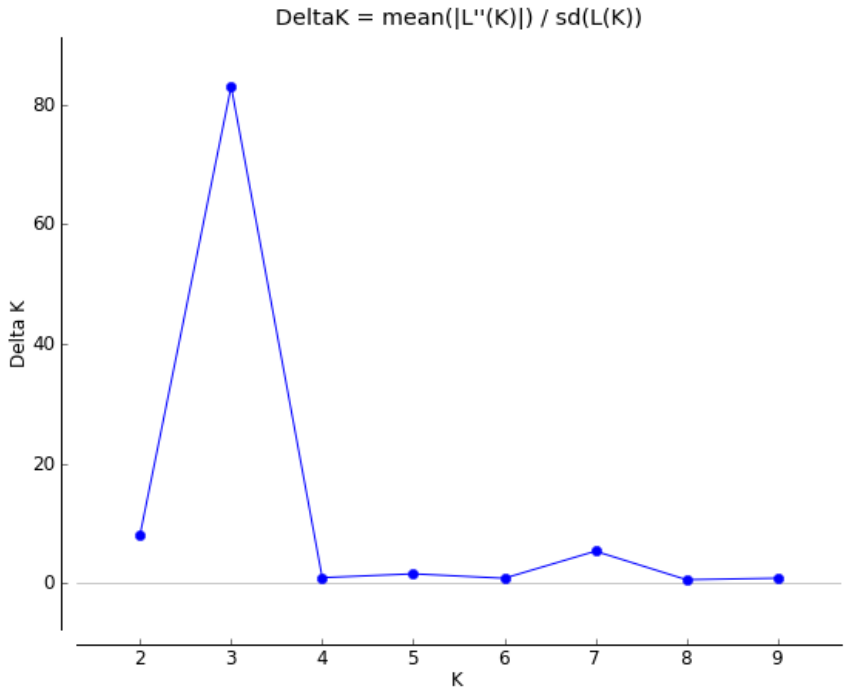

Figure 3.3. Evanno *et al.* (2005) plot detecting the number of *K* groups that best fit the data for *Bixa orellana* var. *urucurana* individuals assessed with 16 SSR loci.

The Central Rondônia group had the highest mean number of alleles per locus (6.876) (Table 3.3), as well as the highest H_E (0.706). The greatest value of H_O was identified in the southern Rondônia group. The group of Roraima and Pará showed lower values for $H_{O,}$ resulting in a higher inbreeding coefficient (0.528) and the lowest apparent outcrossing rate (0.309). This group also showed higher levels of allelic richness (5.558).

The projection of the 170 samples of wild annatto on a two-dimensional plane defined by the first two principal coordinates (Figure 3.4), which explained 57.7% of total variation, showed a tendency to group individuals according to their geographical origin. The first coordinate separated the South Rondônia populations (on the left) from the other populations.

The second coordinate separated the Central Rondônia populations (upper right) from the other populations (lower right). The PCoA results were generally similar to those of the *Structure* analysis, but failed to explain why Corumbiara is a mixture of Groups I and II in the *Structure* analysis.

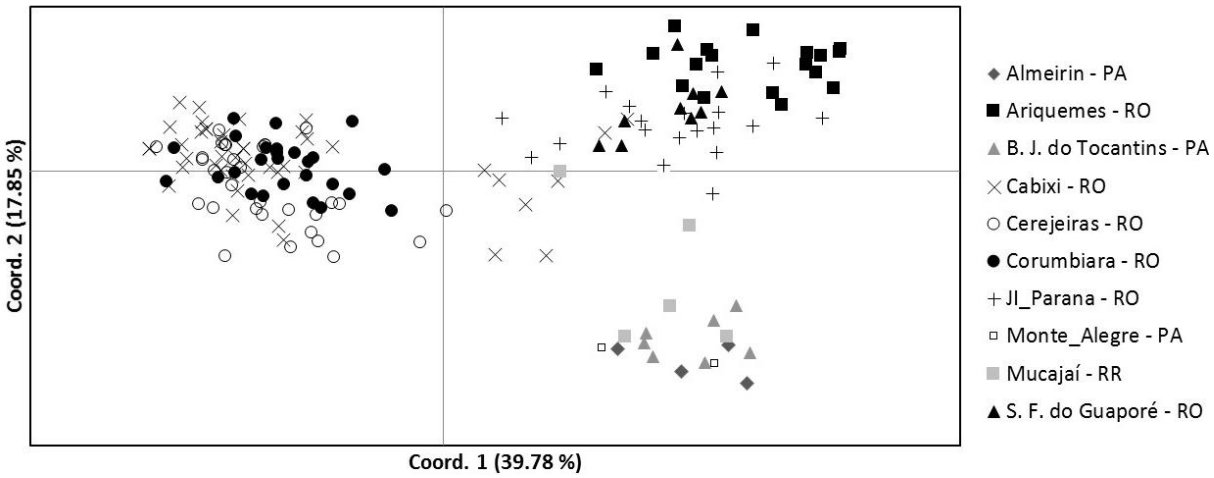

Figure 3.4 Principal coordinate analysis (PCoA) for the 170 samples of 10 wild populations of annatto (*Bixa orellana* var. *urucurana*) collected in Brazilian Amazonia based on 16 SSR loci.

Relationships among populations in the dendrogram (Figure 3.5) generally agreed with *Structure* and PCoA analyses. The relationship among individuals in the dendrogram (Figure 3.6) also agreed with *Structure* and PCoA results. In the individual plants dendrogram (Figure 3.6), the Central Rondônia populations have a slightly greater relationship with the non-Rondônia populations, rather than with the South Rondônia populations, suggesting a difference that may be due to adaptation to the more savanna-like climate of South Rondônia.

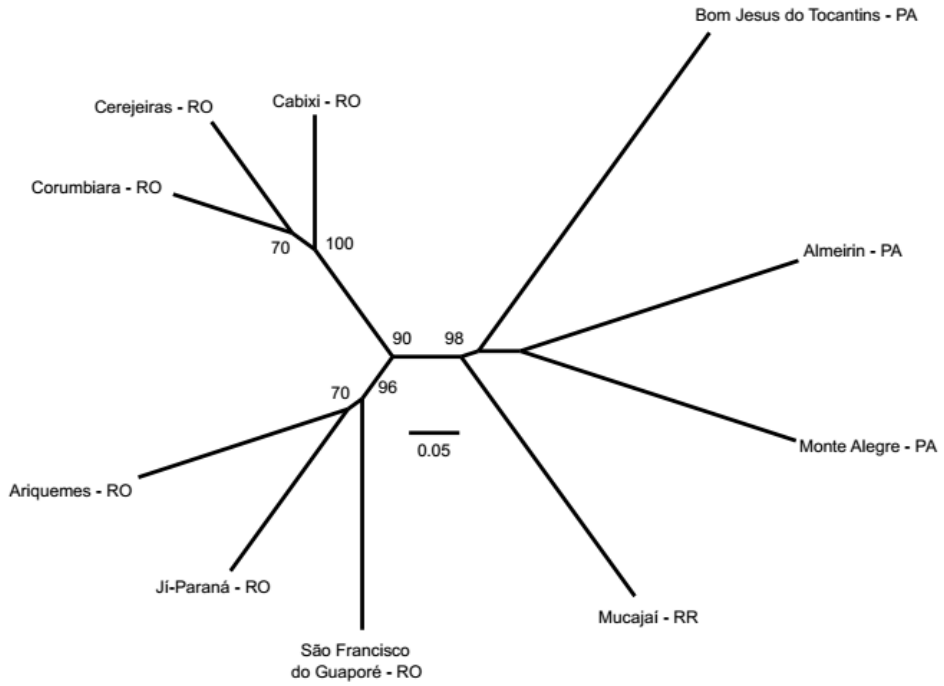

Figure 3.5 Unrooted neighbor-joining dendrogram of 10 *Bixa orellana* var. *urucurana* populations collected in Brazilian Amazonia based on Cavalli-Sforza & Edwards (1967) Chord distance estimated from 16 SSR.

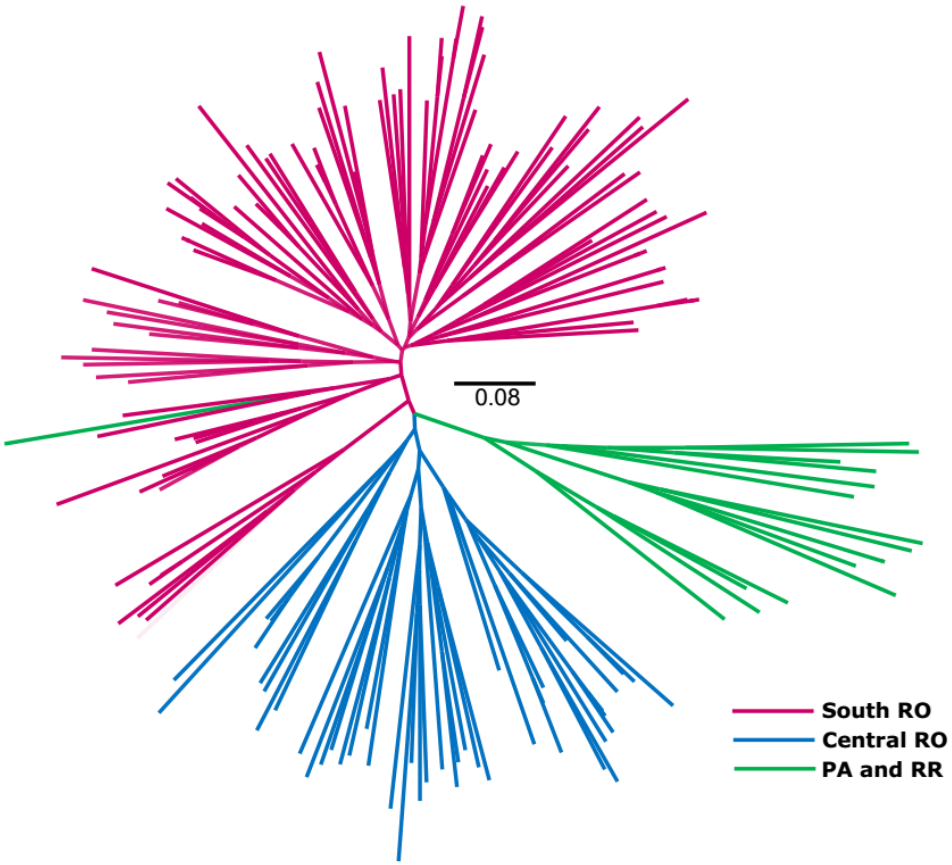

Figure 3.6 Neighbor-joining dendrogram for individuals based on Cavalli-Sforza and Edwards (1967) Chord distance estimated from 16 nuclear microsatellites (SSR) and 170 wild annatto (*Bixa orellana* var. *urucurana*) accessions. Branches are colored according to the *Structure* simulation for $K = 3$.

According to Mantel's test, 74% of the genetic divergence among populations was due to the geographic distances among them. A strong and positive correlation between genetic and geographic distances (r = 0.860, p = 0.003) suggests that genetic differentiation among the 10 populations is due to isolation by distance, not surprising given the long distances between Rondônia and Roraima, especially.

AMOVA revealed that 21% of the genetic variation was among populations, and the remaining 79% was within populations (p = 0.000) when considering the 10 distinct populations. When we performed AMOVA considering three groups according to the most likely *K* of the *Structure* analysis, the genetic variation among populations decreased to 16% (p = 0.000) and still showed that, although most of the diversity is within populations, the variation due to sub-division of the populations is quite significant. The microsatellite data also showed low levels of gene flow among populations (N_m = 0.545). However, when analyzed by the a priori populations, we observed an average gene flow of 1.878 among populations from Rondônia, while the other regions presented lower values (N_m = 0.65). The populations from south Rondônia separately showed a high gene flow among them (N_m = 4.843).

3.3.3. Species distribution modeling and climatic adaptation

The result of PCA analysis using the 10 populations sampled in this study and 174 presence-only records with the 19 bioclimatic variables generated three main components that explained more than 91% of the variation. Graphical representation of climate space associated with the first two PCA axes revealed high climatic differenciation between populations from Rondônia, and Pará and Roraima along the second PCA axis (Figure 3.7). There was also a moderate degree of climatic overlap between Central Rondônia and South Rondônia groups, indicating evidence of incomplete separation between the Rondônia groups according to the bioclimatic variables.

The relative contributions of climatic variables to the PCA axes show that niche differentiation along Components 1 and 2 was driven primarily by precipitation requirements (Table 3.4). Principal component 1 (PC1) represented 71% of the variation and was mostly explained by variable Bio12 (Annual Precipitation). The annual precipitation varied from 1309 mm (on the left) to 3644 mm (on the right). Our sampled populations presented annual precipitations ranging from 1669 mm to 2192 mm. While variable Bio12 contributed

positively, variable Bio15 (Precipitation Seasonality) contributed negatively in the first axis. On the other hand, the second axis explained 20% of the variation and variable Bio19 (Precipitation of Coldest Quarter) was the most informative variable in this axis (Table 3.4), ranging from 87 mm (on the top) to 1388 mm (on the bottom). Our sampled populations ranged from 94 to 917 mm of precipitation in the coldest quarter. Variables Bio16 (Precipitation of Wettest Quarter) and Bio17 (Precipitation of Driest Quarter) also played important roles in the analysis.

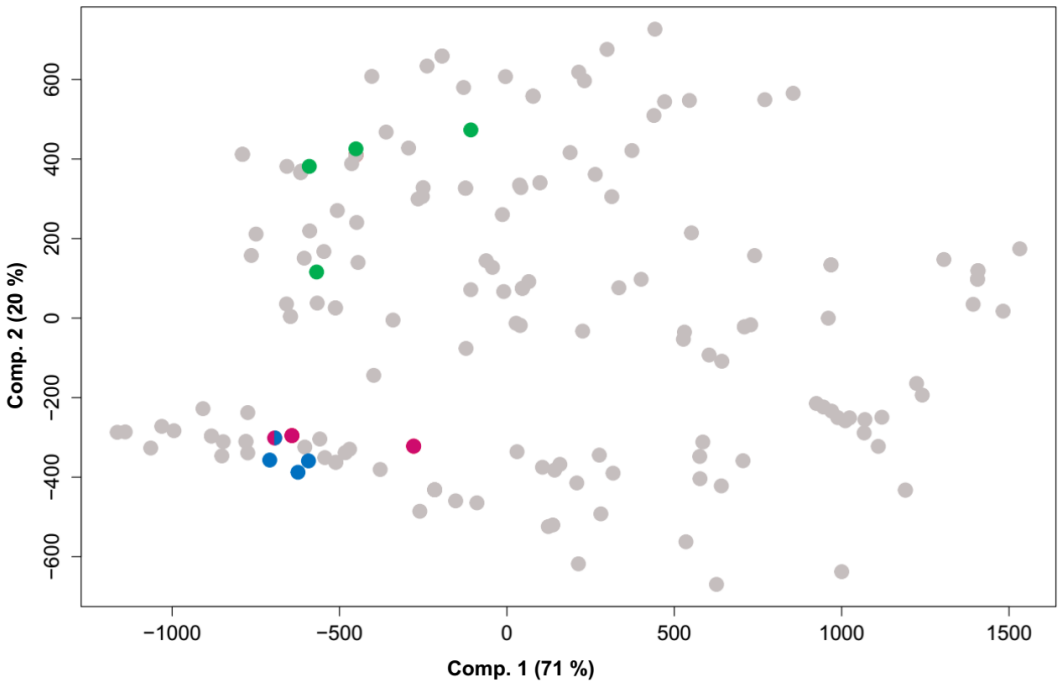

Figure 3.7 Principal component analysis (PCA) performed on 19 bioclimatic variables extracted from the Worldclim database for 10 wild annatto (*Bixa orellana* var. *urucurana*) populations sampled and 174 presence-only records from online databases. Colors are according to the *Structure* analysis: Pink = South Rondônia accessions; Blue = Central Rondônia accessions; Green = accessions from the states of Pará and Roraima and Gray = online databases occourences.

Table 3.4 Factor loadings of Principal component analysis (rotation) on a set of 19 bioclimatic variables retained for their contribution to the model of distribution (higher values in bold characters).

Variable	PC1 (71%)	PC2 (20%)	PC3 (6%)	PC4 (2%)
BIO1 = Annual Mean Temperature	-0.0003	0.0023	0.0002	-0.0003
BIO2 = Mean Diurnal Range	-0.0005	-0.0011	-0.0026	-0.0017
BIO3 = Isothermality (BIO2/BIO7)	0.0032	0.0008	0.0117	0.0080
BIO4 = Temperature Seasonality	-0.0046	0.0028	-0.0271	-0.0600
BIO5 = Max Temperature of	-0.0007	0.0025	-0.0025	-0.0013
BIO6 = Min Temperature of Coldest	0.0006	0.0042	0.0026	0.0022
BIO7 = Temperature Annual Range	-0.0013	-0.0017	-0.0052	-0.0034
BIO8 = Mean Temperature of	-0.0004	0.0012	0.0000	-0.0010
BIO9 = Mean Temperature of Driest	0.0000	0.0037	0.0006	0.0022
BIO10 = Mean Temp. of Warmest	-0.0003	0.0026	-0.0003	-0.0006
BIO11 = Mean Temperature of	-0.0002	0.0024	0.0004	0.0007
BIO12 = Annual Precipitation	**0.8757**	-0.1919	-0.1250	**0.3052**
BIO13 = Precipitation of Wettest	0.0759	0.0567	-0.2527	-0.1002
BIO14 = Precipitation of Driest	0.0581	-0.0527	0.1833	0.0580
BIO15 = Precipitation Seasonality	**-0.0190**	0.0251	-0.0670	-0.0363
BIO16 = Precipitation of Wettest	0.2129	0.1468	**-0.6377**	-0.1877
BIO17 = Precipitation of Driest	0.1946	-0.1584	**0.5789**	0.2289
BIO18 = Precipitation of Warmest	0.2106	**-0.4472**	0.1891	**-0.8459**
BIO19 = Precipitation of Coldest	0.3099	**0.8425**	0.3224	-0.2921

Over 15 replicate runs, the potential distribution of *B. orellana* var. *urucurana* was estimated with a high area-under-the-curve (AUC) value (0.941) implying very low rates of false negative and positive suitability predictions. Figure 3.8 shows the distribution of suitable habitat for *B. orellana* var. *urucurana*. The climate envelope of wild annatto is largely determined by precipitation, and the most importante variables for the model were Bio19 (Precipitation of coldest quarter, 23.5%), Bio13 (Precipitation of wettest month, 12.4%) and Bio12 (Annual precipitation, 11.5%). Temperature seasonality (Bio4, 17.5%) also plays a substantial role in the niche.

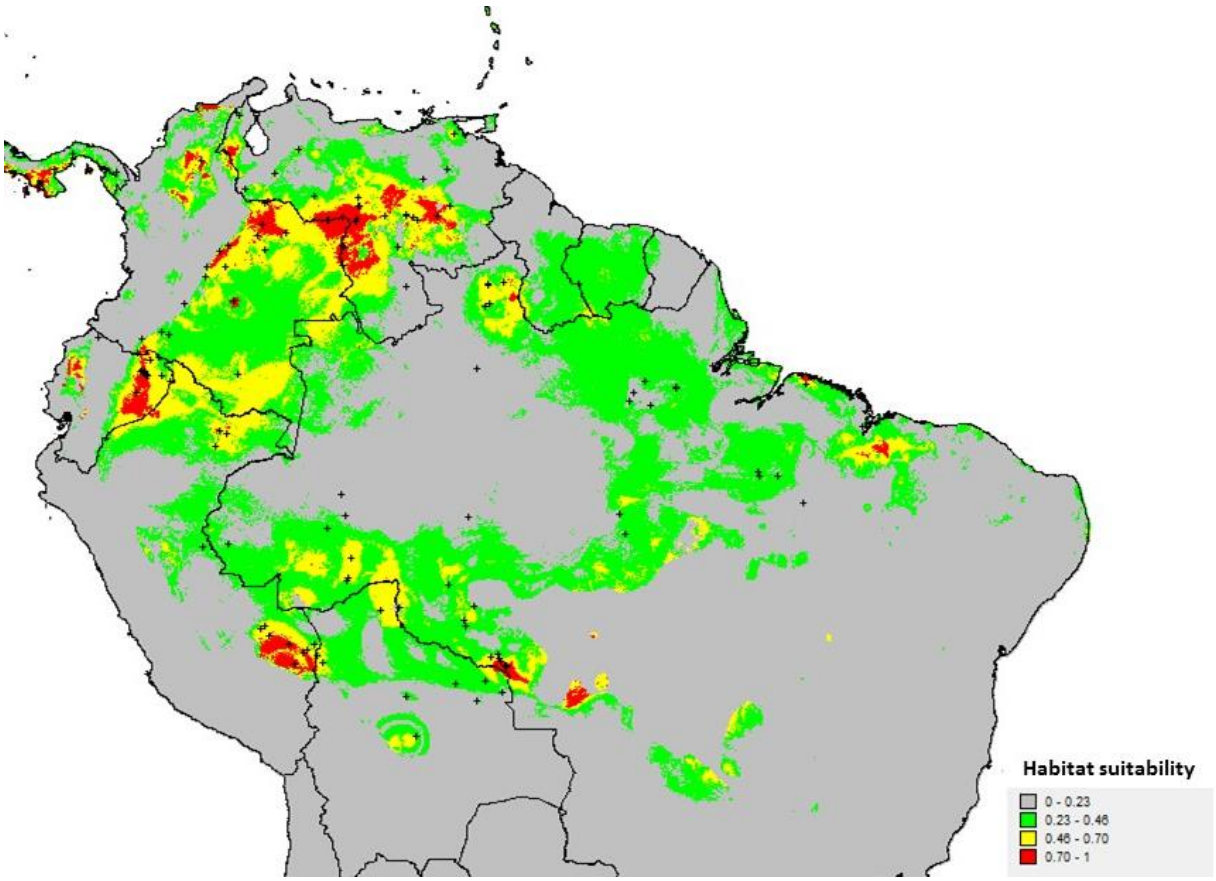

Figure 3.8 Potential distribution as probability of occurrence of *B. orellana* var. *urucurana* simulated by Ecological Niche Modeling. Black crosses are the presence records used for the simulation

3.4. Discussion

3.4.1. Genetic diversity

This is the first genetic study with wild populations of annatto. The levels of heterozygosity averaged over all loci (H_E = 0.520; H_O = 0.385) among the 10 populations were considerably higher than those in a germplasm bank (H_E = 0.470; H_O = 0.170) with 63 cultivated varieties of annatto using the same 16 SSR loci (Chapter 4, Dequigiovanni *et al.*, in press). It is expected that wild populations exhibit higher levels of diversity in relation to cultivated populations, as the latter have passed through a domestication bottleneck (Olsen and Wendel, 2013) and generally further bottlenecks due to distribution and diversification (Meyer and Purugganan, 2013), including different selection pressures for yield (Gepts, 2004).

The high levels of local inbreeding coeficient (*f*) observed in our study, the estimated apparent outcrossing rate ($\hat{t}_a$ = 0.609) and, the apparent selfing rate (0.391), indicate a mixed mating system for annatto. The existence of crosses between related individuals generates

selfing and, therefore, increased homozygosity (Ritland, 2002). These results are in agreement with those obtained by Rivera-Madrid *et al* (2006), who conducted controlled pollinations in annatto accessions in an experimental field in Mexico, suggesting that annatto can tolerate both types of pollination, showing cross-pollination values of 57% and self-fertilization of 31%. Vilares *et al* (1992) also concluded that there is natural selfing in annatto. Valdez-Ojeda *et al* (2010) found high multilocus outcrossing rates (0.75) based on 50 SRAP loci and concluded that *B. orellana* has a mixed mating system.

3.4.2. Genetic structure

Plant populations are not randomly arranged assemblages of genotypes but are structured in space and time. Genetic structure results from the action of migration, mutation, selection, and drift, that operates within the historical and biological context of each plant species (Loveless and Hamrick, 1984). In this study, most of the genetic variability was observed within populations (79%), as shown in the AMOVA results. However, the high value of F_{ST} (0.201) indicates the existence of strong structure among populations. Dick *et al* (2008) reviewed the genetic structure among populations of 42 different tropical species separated by more than 50 km and found high levels of population differentiation (F_{ST} = 0.177). This may be due to the fact that tropical species are experiencing moderate to high levels of inbreeding, as a result of the association between low population density, density-dependent animal pollination, and mixed mating systems, factors that may be restricting gene flow (Dick *et al.*, 2008).

The mean level of gene flow among *B. orellana* var. *urucurana* populations was low in this study (N_m = 0.545), because isolation by distance was high among many of the populations, especially between populations from Rondônia and the ones from Roraima and Pará, but also between those from Roraima and Pará. According to Wright (1931), a migration rate of N_m = 1.0 is theoretically necessary to counteract the effect of genetic drift. In this sense, our results suggest that genetic drift may have been a dominant factor determining the genetic structure of *B. orellana* var. *urucurana* populations. Gene flow among wild annatto populations may occur by seed dispersion along rivers and streams (Moreira *et al.*, 2015), but also through cross-pollination by bees (Costa *et al.*, 2008). All wild annatto populations we found in Amazonia occurred in open forests and anthropogenic landscapes, although never cultivated, and always associated with riparian environments, suggesting that gene flow

among distant populations may have occurred by the transport of fruits, and consequently seeds, along rivers. The differentiation between the Rondônia and the Pará/Roraima populations is also due to the fact that the distances among these populations are very large, in addition to the existence of many barriers that can prevent or slow gene flow. In fact, what is surprising is that the Amazon River did not prove to be a major barrier between the northern populations (Roraima and northern Pará) and the southeastern Pará population until K = 7.

The structure of the Rondônia populations in two groups may occur because the southern populations of this state (Cabixi, Corumbiara and Cerejeiras) are in the Guaporé River basin, while the populations from the center of Rondônia state (Ariquemes and Jí Paraná) are in the Jí-Paraná River basin, so that gene flow by the transport of seeds along the rivers between these populations is infrequent.

The significant correlation values obtained between genetic and geographic distances indicate isolation by distance for the wild annatto populations. In the cluster analysis, based on Cavalli-Sforza and Edwards (1967) Chord distance and the Neighbor-Joining method, as well as in the PCoA and the Bayesian analyses, three genetically distinct and consistent groups were identified. The groups were formed based on the geographical location of the collected populations. Carvalho *et al* (2005), using isozymes, also found that genetic differences in cultivated annatto accessions correspond to geographical differences. However, Medina *et al* (2001) evaluated 36 genotypes of cultivated annatto collected in Venezuela and Brazil, also using isozymes, and found no correlation between geographic and genetic distances. The explanation given by the authors was the anthropogenic influence in this crop´s cultivation. Menezes *et al* (2014) found similar patterns to those obtained in this study when assessing wild cotton (*Gossypium mustelinum* Miers) in the state of Bahia, the only cotton species native to Brazil. The authors found high correlation between the genetic and geographic distances using Mantel's test ($r = 0.87$, $p = 0.05$).

Our results provide relevant information for conservation of annatto germplasm. Wild annatto populations are a source of new alleles, which might be useful to increase the genetic basis of annatto in breeding programs and for conservation strategies. According to Moreira *et al* (2015) and also our field observations, when wild annatto grows near commercial annatto fields, farmers tend to remove the existing wild types, because they naturally cross generating production losses in the progenies. As occurrences of wild annatto are mainly in the peripheries of the Amazonia, and in most cases, in anthropogenic areas (Moreira *et al.*, 2015), these practices may lead to a decrease of wild annatto populations, fragmentation of the native habitat of the species, and overall genetic diversity of the species.

3.4.3. Distribution and climatic adaptation

Temperature and precipitation are considered as major factors in determining species distributions (Wiens, 2011). Our models suggest that precipitation plays a key role in wild annatto's current and potential distribution pattern. In general, favourable habitats are drier or seasonally drier areas, which suggests the species tolerance of drought conditions, even though we observe a wide variation in levels of precipitation in the regions with occurrence of the species. Although temperature variables did not contribute much to the distribution of the species, we observed a wide variation in temperature averages. The large intervals of temperature and precipitation suggest that the species has ample adaptation. Temperature and precipitation have been identifed as majors selective pressures driving plant adaptation because they are very important for plant growth, development, and reproduction (Manel *et al.*, 2010; Wang *et al.*, 2016). Adaptation to new habitats is also a potential plant response to shifts in environmental conditions, which is also crucial in the context of climate change (Reusch and Wood, 2007).

Not surprisingly, the potential distribution coincided approximately with the current occurrence reported in online distribution databases. However, a considerable number of occurences are in very low probability areas, suggesting that this large-scale analysis did a poor job of capturing urucurana's adaptation to riparian conditions in drier climates. The species is mostly confined to the periphery of Amazonia, but also to areas in the drier parts of western Central America. According to Clement *et al* (2010), the periphery of Amazonia appears to be the area where the majority of Amazonian crops were domesticated. The upper Madeira River basin, in southwestern Amazonia, is an important part of the periphery and has been recognized as a probable region of crop origins for some time (Clement *et al.*, 2016). Levis *et al* (2017) also found higher abundance and richness of domesticated species in southwestern Amazonia.

Piperno and Pearsall (1998) also highlighted the importance of the periphery, mainly in extreme northwestern Amazonia and the adjacent Llanos of the Orinoco River basin, the Guiana shield, as well as in southwestern Amazonia, especially the Llanos de Mojos, in Bolivia. The potential distribution map predicted moderately suitable habitat in the Llanos de Mojos. The Llanos de Mojos is a tropical savanna in Bolivian Amazonia, shaped by cycles of drought and flood (Walker, 2008). This grassland environment presents a 2- to 7-month dry season and a total annual rainfall varying between 1,500 and 1,800 mm. The dry season lasts from May through September, when weeks pass without precipitation (Walker, 2008).

Complex societies inhabitated this region at the time of the European conquest, and managed dozens of species, leading Clement (1999) to propose a micro-center of diversity of crop genetics resources in Llanos de Mojos. Also, the only archaeological record of annatto in Amazonia comes from this area, and is dated to 2400 years before present (Erickson, 1995).

The Madre de Dios Basin is also a highly suitable area for *B. orellana* var. *urucurana*. According to Leal and Clavijo (2010), the genus *Bixa* probably originated between the Huallaga-Ucayali River, and the Madre de Dios-Madeira River, along the slopes of the eastern Andes. The Madre de Dios River joins with the Mamore River to become the Madeira River, also an important area for crop domestication (Clement *et al.*, 2016). The Madre de Dios Basin drains an area of approximately 90,000 km along the eastern flank of the Cordillera de los Andes in southeastern Peru, ranging in elevation from 200 m to over 4000 m (Barthem *et al.*, 2003). The vegetation is predominantly evergreen or semi-evergreen forest (Osher and Buol, 1998). It presents a humid tropical climate with annual rainfall varying from 1200 mm to 3300 mm, generally increasing from east to west, and the rainy season occurring from October to April (Osher and Buol, 1998).

We also found highly suitable habitat for wild annatto in the Llanos del Orinoco, in western Venezuela and northeastern Colombia. This is an area of extensive plains, covered mainly by savanna vegetation. This ecoregion has a strongly seasonal climate, with a single dry season extending between November and May, and a single rainy season between May and October. The temperature prevailing in these tropical American lowlands is macrothermic, with mean anual temperatures ranging from 26ºC to 28ºC and monthly average maximum temperatures between 34ºC and 37ºC. The rainfall of the Llanos region shows a regime characterized by very pronounced differences during the months of the year, with annual rainfall amount ranging from 850 mm to 1800 mm (Stefano *et al.*, 2006).

The areas of the Llanos de Orinoco, Llanos de Mojos, Madre de Dios and also South of Rondônia have very similar climatic characteristics, which make all of them areas suitable for the occurrence of the species. These are areas that present drier or seasonally drier areas and are located in the peripheries of the Amazon, consistent with the favorable areas identified by Moreira *el al* (2015). In Rondônia, our sampled populations in the savannas of South Rondônia are in an area of very high probability in the potential distribution map, while Central Rondônia populations are in an area with much less probability, and this may suggest differential adaptation.

On the other hand, an interesting result of the ENM model was the high probability area in eastern Ecuador. The eastern lowlands in Equador experience abundant rainfall,

sometimes exceeding 5,000 mm per year and mean temperatures ranging from 25°C to 28°C. These findings also suggest adaptation of *B. orellana* var. *ururucana* to different niches.

3.5. Conclusion

The microsatellites loci used in this study revealed high levels of genetic diversity among populations of wild annatto and this diversity is highly structured according to the geographic origin of populations. Wild annatto appears to have a mixed mating system, which may contribute to the patterns of genetic structure observed. Our map of the potential distribution of the species allowed the identification of other potential areas of occurrence in Amazonia and in northern South America. New plant collections will add to a better understanding of the genetic diversity and structure of wild annatto, as well as the understanding of the crop's domestication from these wild populations.

REFERENCES

Akshatha, V., Giridhar, P. and Ravishankar, G.A. (2011) Morphological diversity in *Bixa orellana* L. and variations in annatto pigment yield. *J. Hortic. Sci. Biotechnol.* **86**, 319–324.

Anon (2013) The Plant List Version 1.1. *Publ. internet*.

Arce, J. (1999) El achiote *Bixa orellana* L. cultivo promisorio para el trópico. In Earth., p. 149.

Baer, D.. (1976) *Systematics of the genus Bixa and geography of the cultivated annatto*. University of California - Los Angeles.

Barthem, R., Goulding, M., Forsberg, B., Canas, C. and Ortega, H. (2003) Aquatic ecology of the Rio Madre de Dios. Scientific bases for Andes Amazon Headwaters (ACCA), ed, Lima, Peru: Gráfica Biblos.

Carvalho, J.F.R.P. de, Robinson, I.P. and Alfenas, A.C. (2005) Isozymic variability in a Brazilian collection of annatto (*Bixa orellana* L.). *Pesqui. Agropecuária Bras.* **40**, 653–660.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967) Phylogenetic analysis models and estimation procedures. *Am. J. Hum. Genet.* **19**, 233–57.

Clement, C.R. (1999) 1492 and the loss of amazonian crop genetic resources. I. The relation between domestication and human population decline. *Econ. Bot.* **53**, 188–202.

Clement, C.R., Cristo-Araújo, M. de, D'Eeckenbrugge, G.C., Pereira, A.A. and Picanço-Rodrigues, D. (2010) Origin and domestication of native Amazonian crops. *Diversity* **2**, 72–106.

Clement, C.R., Rodrigues, D.P., Alves-Pereira, A., Mühlen, G.S., Cristo-Araújo, M. De, Moreira, P.A., Lins, J. and Reis, V.M. (2016) Crop domestication in the upper Madeira River basin. *Bol. do Mus. Para. Emilio GoeldiCiencias Humanas* **11**, 193–205.

Coppens d'Eeckenbrugge, G. and Lacape, J.M. (2014) Distribution and differentiation of wild, feral, and cultivated populations of perennial upland cotton (*Gossypium hirsutum* L.) in Mesoamerica and the Caribbean Zhang, X., ed. *PLoS One* **9**, e107458.

Costa, A., Guimarães-Dias, F. and Pérez-Maluf, R. (2008) Abelhas (*Hymenoptera: Apoidea*) visitantes das flores de urucum em Vitória da Conquista, BA. *Ciência Rural* **38**, 534–537.

Dequigiovanni, G., Ramos, S.L.F., Zucchi, M.I., Bajay, M.M., Pinheiro, J.B., Fabri, E.G., Bressan, E.A. and Veasey, E.A. (2014) Isolation and characterization of microsatellite loci for *Bixa orellana*, an important source of natural dyes. *Genet. Mol. Res.* **13**.

Dick, C.W., Hardy, O.J., Jones, F.A. and Petit, R.J. (2008) Spatial scales of pollen and seed-mediated gene flow in tropical rain forest trees. *Trop. Plant Biol.* **1**, 20–33.

Dieringer, D. and Schlötterer, C. (2003) Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* **3**, 167–169.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of Plant DNA from fresh tissue. *Focus (Madison).* **12**, 13–15.

Dray, S. and Dufour, A.-B. (2007) The ade4 Package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20.

Ducke, A. (1946) Plantas de cultura pré-colombiana na Amazônia brasileira: notas sobre as espécies ou formas espontâneas que supostamente lhes teriam dado origem. *Bol. do Inst. Agronômico do Norte* **8**, 1–24.

Elith, J., H. Graham, C., P. Anderson, R., *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography (Cop.).* **29**, 129–151.

Erickson, C. (1995) Archaeological methods for the study of ancient landscapes of the Llanos de Mojos in the Bolivian Amazon. In Archaeology in the lowland American tropics: Current analytical methods and applications. (Stahl, P., ed), pp. 66–95. Cambridge: Cambridge University Press.

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**, 2611–20.

Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6.

Fielding, A.H. and Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence / absence models. *Environ. Conserv.* **24**, 38–49.

Gepts, P. (2004) Crop domestication as a long term selection experiment, Oxford, UK: John Wiley & Sons, Inc.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978.

Jombart, T. and Ahmed, I. (2011) Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071.

Kamvar, Z.N., Tabima, J.F. and Grünwald, N.J. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281.

Keenan, K., McGinnity, P., Cross, T.F., Crozier, W.W. and Prodöhl, P.A. (2013) diveRsity : An R package for the estimation and exploration of population genetics parameters and their associated errors O'Hara, R.B., ed. *Methods Ecol. Evol.* **4**, 782–788.

Kuntze, C. (1925) Bixaceae. In Die natürlichen Pflanzenfamilien. (Engler, A. and Prantl, K., eds), p. 315. Leipzig: Engelmann.

Leal, F. and Clavijo, C.M. (2010) Acerca de la história, taxonomia, botánica y usos de *Bixa orellana* L. *Rev. Unell. Cienc. Tec* **1**, 78–86.

Levis, C., Costa, F.R.C., Bongers, F., *et al.* (2017) Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. *Science (80).* **355**, 925–931.

Loveless, M.D. and Hamrick, J.L. (1984) Ecological determinants of genetic structure in plant populations. *Annu. Rev. Ecol. Syst.* **15**, 65–95.

Manel, S., Poncet, B.N., Legendre, P., Gugerli, F. and Holderegger, R. (2010) Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Mol. Ecol.* **19**, 3824–3835.

Medina, A.M., Michelangeli, C.C., Ramis, C.M. and Díaz, A.J. (2001) Caracterización morfológica de frutos de onoto (*Bixa orellana* L.) y su correspondencia con patrones de proteínas e isoenzimas. *Acta Científica Venez.* **52**, 14–23.

Menezes, I.P.P. de, Gaiotto, F.A., Hoffmann, L.V., Ciampi, A.Y. and Barroso, P.A.V. (2014) Genetic diversity and structure of natural populations of *Gossypium mustelinum*, a wild relative of cotton, in the basin of the De Contas River in Bahia, Brazil. *Genetica* **142**, 99–108.

Meyer, R.S., DuVal, A.E. and Jensen, H.R. (2012) Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.* **196**, 29–48.

Meyer, R.S. and Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852.

Moreira, P.A., Lins, J., Dequigiovanni, G., Veasey, E.A. and Clement, C.R. (2015) The domestication of Annatto (*Bixa orellana*) from *Bixa urucurana* in Amazonia. *Econ. Bot.* **69**, 127–135.

Mousadik, A. El and Petit, R.J. (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theor. Appl. Genet.* **92**, 832–839.

Nisar, N., Li, L., Lu, S., Khin, N.C. and Pogson, B.J. (2015) Carotenoid metabolism in plants. *Mol. Plant* **8**, 68–82.

Olsen, K.M. and Wendel, J.F. (2013) Crop plants as models for understanding plant adaptation and diversification. *Front. Plant Sci.* **4**, 290.

Osher, L.. and Buol, S.. (1998) Relationship of soil properties to parent material and landscape position in eastern Madre de Dios, Peru. *Geoderma* **83**, 143–166.

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.

Peterson, A.T. (2003) Predicting the geography of species' invasions via ecological niche modeling. *Q. Rev. Biol.* **78**, 419–33.

Phillips, S.J., Anderson, R.P. and Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecol. Modell.* **190**, 231–259.

Piperno, D.R. and Pearsall, D.M. (1998) The origins of agriculture in the lowland neotropics, Academic Press.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

R Core Team (2015) R: A language and environment for statistical computing.

Reif, J.C., Melchinger, A.E. and Frisch, M. (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Sci.* **45**, 1.

Reusch, T.B.H. and Wood, T.E. (2007) Molecular ecology of global change. *Mol. Ecol.* **16**, 3973–92.

Ritland, K. (2002) Extensions of models for the estimation of mating systems using n independent loci. *Heredity (Edinb).* **88**, 221–8.

Rivera-Madrid, R., Escobedo-GM, R.M.M., Balam-Galera, E., Vera-Ku, M. and Harries, H. (2006) Preliminary studies toward genetic improvement of annatto (*Bixa orellana* L.). *Sci. Hortic. (Amsterdam).* **109**, 165–172.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25.

Sandy-Cuen, P.M. and Becerra, R. (2003) Manejo campesino de recursos naturales. El achiote. *BioDiversitas* **7**, 7–11.

Schuelke, M. (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18**, 233–234.

Schultes, R. (1984) Amazonian cultigens and their northward and westward migrations in pre-Columbian times. In Pre-Columbian Plant Migration. Papers of the Peabody Museum of Archaeology and Ethnology. (Stone, D., ed), pp. 19–38. Cambridge: Harvard University.

Slatkin, M. and Barton, N.H. (1989) A comparison of three indirect methods for estimating average levels of gene flow. *Evolution (N. Y).* **43**, 1349.

Stefano, R., Aymard, G., Riina, R. and Huber, O. (2006) Flora and vegetation of the Venezuelan Llanos: a review. In Neotropical savannas and seasonally dry forests plant diversity, biogeography, and conservation. (Pennington, R., Lewis, G., and Ratter, J., eds), pp. 96–118. New York.

Valdez-Ojeda, R., Hern??ndez-Stefanoni, J.L., Aguilar-Espinosa, M., Rivera-Madrid, R., Ortiz, R. and Quiros, C.F. (2008) Assessing morphological and genetic variation in Annatto (*Bixa orellana* L.) by sequence-related amplified polymorphism and cluster analysis. *HortScience* **43**, 2013–2017.

Valdez-Ojeda, R., Quiros, C.F., Lourdes Aguilar-Espinosa, M. de, Rivera-Madrid, R., Aguilar-Espinosa, M. de L. and Rivera-Madrid, R. (2010) Outcrossing rates in annatto determined by sequence-related amplified polymorphism. *Agron. J.* **102**, 1340–1345.

Vencovski, R. (1994) Variance of an estimative of the outcrossing rate. *Rev. Bras. Genética* **17**, 349–351.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. (2016) Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **39**, 312–328.

Vilares, A.S., São José, A.R., Rebouças, T.N.H. and Souza, I.V.B. (1992) Estudo da biologia floral de urucuzeiro (*Bixa orellana* L.). *Rev. Bras. Corantes Nat.* **1**, 101–105.

Walker, J.H. (2008) The Llanos de Mojos. In The Handbook of South American Archaeology., pp. 927–939. New York, NY: Springer New York.

Wang, T., Wang, Z., Xia, F. and Su, Y. (2016) Local adaptation to temperature and precipitation in naturally fragmented populations of *Cephalotaxus oliveri*, an endangered conifer endemic to China. *Sci. Rep.* **6**, 25031.

Wiens, J.J. (2011) The niche, biogeography and species interactions. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366**, 2336–2350.

Wright, S. (1931) Evolution in Mendelian Populations. *Genetics* **16**, 97–159.

Wright, S. (1965) The interpretation of population structure by f-statistics with special regard to systems of mating. *Evolution (N. Y).* **19**, 395.

4. GENETIC DIVERSITY AND STRUCTURE IN A MAJOR BRAZILIAN ANNATTO (*BIXA ORELLANA*) GERMPLASM BANK REVEALED BY MICROSATELLITES AND PHYTOCHEMICAL COMPOUNDS

Abstract

Annatto (*Bixa orellana* L.) is a tropical crop indigenous to the Americas, probably Amazonia. Annatto is commercially valuable in the food and cosmetics industries as a natural dye to be used instead of synthetic dyes. In addition, annatto contains other important substances for human health, such as geranylgeraniol, tocotrienols and other carotenoids. The aim of the present study was to evaluate the genetic diversity of 63 accessions from the annatto germplasm bank at the Agronomic Institute (IAC), São Paulo, Brazil, using four phytochemical compounds and 16 microsatellite markers. Significant variation was observed for the phytochemical compounds, ranging from 2 to 7.31 g (100 g dry matter - DM)$^{-1}$ for bixin, 2.14 to 7.11 g (100 g DM)$^{-1}$ for lipids, 0.25 to 1.05 g (100 g DM)$^{-1}$ for tocotrienols, and 0.49 to 2.61 g (100 g DM)$^{-1}$ for geranylgeraniol contents. A total of 73 alleles was observed in the molecular characterization with 16 microsatellite loci. We found higher expected heterozygosity than observed heterozygosity for all loci, indicating strong deficits of heterozygotes. For both molecular and phytochemical compounds, cluster and PCoA analyses tended to separate the accessions from Rondônia, northern Brazil, with only a few exceptions, from the Southwestern accessions. The same two groups were found in the Bayesian analysis with molecular data. Rondônia accessions showed higher values for all the phytochemical compounds and higher levels of genetic diversity. Some accessions presented bixin levels well above the average and are promising materials to be used in genetic improvement programs.

Keywords: Bixin; Lipids; Tocotrienols; Geranylgeraniol; SSR; Diversity

4.1. Introduction

Annatto (*Bixa orellana* L.) is a crop native to continental tropical America (Arce 1999), probably Amazonia (Sandy-Cuen and Becerra 2003; Clement *et al.* 2010), with the wild variety (*B. orellana* var. *urucurana*) naturally occurring in open forests always associated with riparian environments, presenting a wide geographical distribution in Amazonia, northern South America and Central America. The domesticated types have been cultivated since ancient times and grow from Mexico to Argentina and throughout the Caribbean Islands in the Americas. The plant is also naturalized and cultivated in tropical and subtropical areas of other continents (Franco *et al.*, 2008; Leal and de Clavijo, 2012; Moreira *et al.*, 2015).

The only recent systematic treatment of *Bixa* (Baer, 1976) accepted *B. orellana* L. Sp. Pl. (1753) 512 as the correct specific name, reducing a number of previous names to

synonymy: *B. mericana* oir. Encycl. 6 (1804) 229; *B. rleana* oronha. Verh. Batav. Genootsch. Kunsten 5(4) (1790) 8; *B. atangensis* Delpierre. Taxon 19 (1970) 304; *B. odorata* Ruiz & Pav. ex G. Don. Gen. Hist. 1 (1831) 295; *B. Orellana* var. *leiocarpa* (Kuntze) Standl. & L.O.Williams. Fieldiana Bot. 29 (1961) 358; *B. orellana* f. *leiocarpa* (Kuntze) J.F.Macbr. Publ. Field Mus. Nat. Hist., Bot. Ser. 14(4) (1941) 11; *B. purpurea* Sweet. Hort. Brit. 33 (1826); *B. tinctaria* Salisb. Prodr. Stirp. Chap. Allerton (1796) 369; *B. upatensis* Ram. Goyena. Fl. Nicarag. 1 (1909) 217; *Orellana americana* (Poir.) Kuntze. Revis. Gen. Pl. 1 (1891) 44; *O. americana* var. *leiocarpa* Kuntze. Revis. Gen. Pl. 1 (1891) 45; *O. orellana* (L.) Kuntze. Revis. Gen. Pl. 3(2) (1898) 9. This treatment is accepted by the Flora of Brazil.

Bixa orellana is a small tree or shrub measuring from 3 to 8 meters in height. It presents simple and glabrous (when adult) leaves, measuring on average 8 cm in length and 4 cm in width, with long petioles, arranged alternately along the branches (Franco *et al.*, 2008; Leal and Clavijo, 2010, 2012). The flowers are large, with colors varying from white to several shades of pink and lilac. Flowers are hermaphrodite, with five sepals, appearing at the end of the branches, forming fascicles. The fruit is a dehiscent ovoid capsule, with two or three carpels, covered with flexible spines when juvenile. It may be reddish, greenish or brown, measuring three to five centimeters long, containing many black seeds covered by a reddish waxy aril with characteristic odor. The fruits are arranged in bunches with up to 17 units. A well-developed fruit can contain 40 to 60 seeds.

Brazil is the major producer of annatto, followed by Peru and Kenya, but the crop is also found in the Dominican Republic, Colombia, Jamaica, Costa Rica, Suriname and other countries in Asia (Akshatha *et al.* 2011). In Brazil, annatto is grown in several regions, with the states of São Paulo, Rondônia and Pará the largest producers. The largest companies that process annatto are located mainly around the city of São Paulo and in the metropolitan region of Campinas in the state of São Paulo (Fabri 2015). Annatto is commercially valuable due to the application of its pigments in the food and cosmetics industries. The principal pigment in annatto extract is bixin, which is found in the resinous coating of the seed itself (Nisar *et al.* 2015). This extract is noteworthy because of its lack of toxicity, its intense coloring capacity and its range of colors, comprising red, orange and yellow hues (Alves *et al.* 2006). Recently, this crop has acquired further importance for containing other important substances for human health, such as geranylgeraniol, tocotrienols and other carotenoids (Albuquerque and Meireles 2012), including the potential treatment of the important tropical and subtropical disease leishmaniasis, caused by *Leishmania braziliensis* and *L. amazonensis*, using annatto's essential oil (Monzote *et al.* 2006; Lopes *et al.* 2012). Annatto seed extracts also showed

insect repellent properties, including protection against *Aedes aegypti*, the insect vector of yellow fever, dengue, chikungunya, zika and other diseases (Giorgi *et al.* 2013).

The use of annatto as a colorant is not new. The Aztecs used annatto extract as a dye for textiles, body paint (such as in lipsticks), and as a food colorant in the drink *cacahuatl* (Giuliano *et al.* 2003). The annatto seed is characterized by the presence of an aril on its surface that contains multiple substances besides the characteristic red pigment. This aril represents about 5-10 % of the seed weight, of which 30 % is the carotenoid bixin. The remaining 70 % is composed of carbohydrates (32 %), lipids (30 %), humidity (3.5 %), protein (2.5 %) and ash (2.0 %) (Carvalho *et al.* 1991). Geranylgeraniol is a naturally occurring linear diterpene soluble in organic solvents, such as chloroform, acetone and alcohol. Geranylgeraniol is an important intermediate of vitamin K, tocopherols and many hormones, and in carotenoid biosynthesis. The presence of geranylgeraniol in annatto seeds was initially described by Craveiro *et al.* (1989), followed by Jondiko and Pattenden (1989), which established the concentration of approximately 1 g (100 g)$^{-1}$ of this metabolite in annatto seeds. Tocotrienols are substances that exhibit strong antioxidant activity and are commonly known as vitamin E. According to Tan and Foley (2002), annatto is one of the few plants containing tocotrienols in a much higher proportion than tocopherols, generating great interest from the pharmaceutical industry. Although found in low amounts in the human diet, tocotrienols are abundant in rice (*Oryza sativa* L.), palm oil (*Elaeis guineensis* Jacq.), and annatto. Tocotrienols are neuro-protective, anti-cancer and cholesterol lowering (Sen *et al.* 2007). Frega *et al.* (1998) described the presence of tocotrienols in annatto seeds at concentrations of 0.14 g (100 g)$^{-1}$. The concentration of carotenoids in annatto seeds usually ranges from 3.12 g (100 g DM)$^{-1}$ to 6.26 g (100 g DM)$^{-1}$. The most important carotenoid in annatto seed is cis-bixin, a monomethyl ester of dicarboxylic acid corresponding to more than 80 % of the total carotenoid content of annatto seeds (Carvalho *et al.* 1993; Auttachoat *et al.* 2011). Despite its growing economic importance, the conservation and manipulation of annatto genetic resources represents the main and most difficult goal for this crop. Considered as the center of origin of annatto, Brazil hosts the greatest diversity of this species. Therefore, characterizing this diversity is a priority in order to promote the conservation of genotypes, as well as to provide information for breeding programs (Rodrigues 1995).

Microsatellite markers or simple sequence repeats (SSR) show high polymorphism, co-dominance and multiallelism. Besides presenting highly reproducible results, this marker is widely distributed in the nuclear genome, making it a useful tool for assessing the genetic diversity and structure of plant populations (Kalia *et al.* 2011; Vieira *et al.* 2016).

Microsatellite primers were recently developed by Dequigiovanni *et al.* (2014) and this is the first report using this marker to analyze the genetic diversity of annatto accessions in Brazil. Although annatto is an important source of natural dye, there is almost no research addressing the genetic variability within this species and investigations on how this variability is distributed in Brazil. In this context, the aim of this study was to characterize the genetic diversity and structure of accessions from the annatto germplasm bank at the Agronomic Institute (IAC), Campinas, São Paulo, Brazil, using microsatellite markers and four phytochemical compounds.

4.2. Material and methods

4.2.1. Plant material

Sixty-three accessions from the annatto Germplasm Bank of IAC, maintained at the Polo Regional Centro Norte in Pindorama, São Paulo, were evaluated in this study. The accessions originated from the Brazilian Central-West, Southeast, and North regions, as well as one accession from Peru and 20 accessions of unknown origin (Figure 4.1; Table 4.1). The collection is maintained in the field, with plots of six half-sib plants, and young leaves from one plant from each accession were collected for the molecular analysis, conducted at the Genetics Department of Luiz de Queiroz College of Agriculture, University of São Paulo.

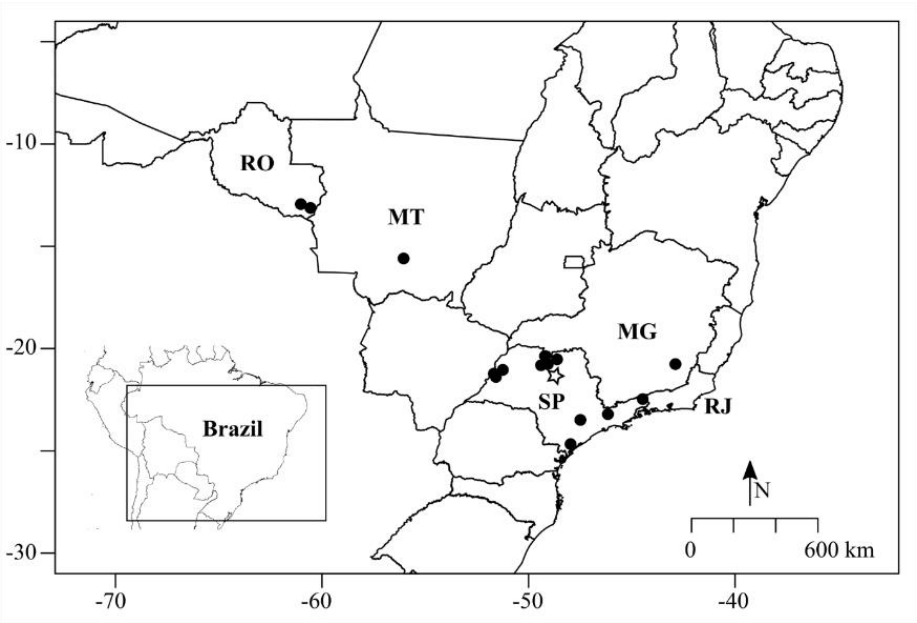

Figure 4.1 Map of Brazil indicating the origins of the annatto (*Bixa orellana*) accessions maintained in the IAC Germplasm Bank. Green circles represent the origin of accessions, while the red star indicates the location of the Germplasm Bank in Pindorama, SP. MG = Minas Gerais, MT = Mato Grosso, RJ = Rio de Janeiro, RO = Rondônia and SP = São Paulo

Table 4.1 Mean values of moisture, lipids, bixin, tocotrienols and geranylgeraniol [g (100 g DM)$^{-1}$] in annatto (*Bixa orellana*) seeds of 63 accessions from the Instituto Agronômico germplasm bank in Campinas, São Paulo, Brazil

ID	Origin	Moisture	Lipids	Bixin	Tocotrienols	Geranylgeraniol
1	Cuiabá-MT	12.94	2.24	3.05	0.41	0.69
2	Cuiabá-MT	8.80	2.87	3.52	0.76	1.13
3	Vale do Ribeira-SP	8.75	2.85	3.57	0.77	1.08
4	Vicosa-MG	10.44	2.63	3.88	0.49	0.99
5	Rezende-RJ	11.36	3.02	2.82	0.80	0.82
6	Sorocaba-SP	9.69	3.43	2.89	0.65	0.96
7	Igaratá-SP	9.96	3.08	2.71	0.39	1.13
8	Tupi Paulista-SP	10.00	3.24	2.60	0.64	1.78
9	Unknown	10.37	3.40	2.75	0.38	1.50
10	Guaraci-SP	8.91	2.97	3.76	0.61	0.98
11	São José do Rio Preto-SP	9.71	3.43	2.82	0.50	1.19
12	Unknown	10.34	3.41	3.18	0.54	1.18
13	Unknown	11.00	3.45	2.46	0.78	1.70
14	Icem-SP	10.14	2.94	3.22	0.74	1.12
15	Unknown	11.27	3.43	2.67	0.59	1.26
16	Minas Gerais-MG	10.88	3.00	4.05	0.48	1.49
17	Barretos-SP	9.23	3.41	2.80	0.51	1.70
18	Unknown	10.61	3.45	2.01	0.46	1.57
19	Unknown	5.78	2.81	2.77	0.43	1.48
20	Icem-SP	9.82	3.38	2.90	0.73	0.93
21	Olímpia-SP	8.73	2.99	2.28	0.65	0.85
22	Peru	11.40	2.48	2.15	0.25	0.56
23	Unknown	12.37	4.66	4.78	1.05	1.10
24	Unknown	9.78	5.00	7.31	1.02	1.57
25	Unknown	10.74	2.55	3.10	0.61	0.63
26	Unknown	10.71	3.36	3.79	0.92	0.82
27	Unknown	8.39	2.14	2.73	0.82	0.49
28	Unknown	10.51	3.42	4.25	0.72	0.77
29	Unknown	8.57	3.28	2.89	0.59	0.88
30	Unknown	11.43	3.16	4.68	0.70	0.76
31	Unknown	10.80	4.40	4.49	1.00	0.62
32	Unknown	11.88	4.01	4.71	0.55	1.21
33	S. João do Pau D'alho	9.60	3.37	3.31	0.93	0.78
34	Unknown	10.11	3.53	3.13	0.43	1.28
35	Unknown	9.46	2.64	3.25	0.62	1.05
36	Unknown	10.61	3.33	2.00	0.47	1.03
37	Unknown	10.31	3.40	3.88	0.82	0.99
38	Corumbiara-RO	12.54	6.91	6.63	1.04	1.82
39	Corumbiara-RO	10.25	4.38	6.56	1.03	1.63
40	Colorado do Oeste-RO	12.67	6.50	4.57	0.94	1.53
41	Colorado do Oeste-RO	12.54	6.43	4.43	0.97	1.78
42	Colorado do Oeste-RO	12.66	3.29	4.02	0.71	1.15
43	Corumbiara-RO	11.32	4.47	4.32	0.63	0.82
44	Corumbiara-RO	12.37	5.53	2.56	0.74	2.45
45	Corumbiara-RO	14.32	4.68	3.35	0.50	1.51
46	Corumbiara-RO	11.33	4.41	3.85	0.65	1.91
47	Corumbiara-RO	10.12	4.29	4.54	1.05	1.46
48	Colorado do Oeste-RO	12.59	4.15	3.39	0.64	0.88
49	Colorado do Oeste-RO	12.27	3.88	4.66	0.66	1.40
50	Colorado do Oeste-RO	9.79	3.11	4.39	0.70	1.11
51	Colorado do Oeste-RO	10.32	3.81	3.67	0.65	1.98
52	Corumbiara-RO	9.92	2.97	4.61	0.66	0.97
53	Colorado do Oeste-RO	8.52	3.39	4.91	0.42	1.30
54	Colorado do Oeste-RO	10.95	3.62	2.77	0.54	1.06
55	Colorado do Oeste-RO	10.87	3.51	4.66	0.61	1.45
56	Colorado do Oeste-RO	12.71	4.90	3.41	0.66	2.61
57	Corumbiara-RO	12.75	4.92	2.68	0.52	2.15
58	Colorado do Oeste-RO	13.14	4.06	4.00	0.97	1.24
59	Colorado do Oeste-RO	11.42	3.17	4.05	0.64	1.49
60	Corumbiara-RO	12.10	5.67	3.99	0.70	1.52
61	Colorado do Oeste-RO	13.98	6.97	5.08	0.64	1.94
62	Colorado do Oeste-RO	12.79	7.11	5.13	0.65	2.04
63	Colorado do Oeste-RO	12.34	2.61	3.08	0.38	0.84
	Overall Mean	10.83	3.79	3.69	0.67	1.27

4.2.2. Phytochemical analysis

For the phytochemical analysis, annatto seeds from the harvest of 2011 were used. Fruits from the selected plants were harvested and dried in the sun. After drying, the seeds were separated from the fruits (capsules) manually. From 100 to 200 g seeds were placed in plastic pots, which were identified and sent to the laboratory at the Institute of Food Technology (ITAL), Campinas, SP, where the analyses were carried out. In the laboratory, the seeds were then transferred to glass containers, where they were kept away from light and refrigerated until the analyses began.

Moisture determination was based on the method described by AOAC (Horwitz, 2005). The determination of lipids was conducted based on the method 2006.06 described by AOAC (Horwitz, 2005) using hexane. The analytical method for the determination of total carotenoids expressed as bixin was based on the saponification of bixin, dilution with potassium hydroxide solution and spectrophotometric quantification, as described by Carvalho *et al.* (2010).

The analytical method used for the determination of tocotrienols and geranylgeraniol was based on saponification with potassium hydroxide solution, extraction of the unsaponifiable fraction with ethyl ether, and transfer of analytes to n-hexane. The analytical method for the determination of tocotrienols was based on the methodology described by Panfili *et al.* (2003). The tocotrienols analysis was performed in a Prominence LC-20A liquid chromatograph coupled to a fluorescence detector RF-10AXL (Shimadzu, Tokyo), using the excitation wavelength of 292 nm and emission of 326 nm. The analytes were resolved on a normal phase column of LiChrospher Si60 (12,5 cm long x 4 mm d.i. and 5 µm particle diameter; Merck, Darmstadt, Germany), having as the mobile phase n-hexane:ethyl acetate:acetic acid (97.6:1.8:0.6, v/v/v), in an isocratic system. The determination of geranylgeraniol was based on the methodology described by Zanh *et al.* (2000). An Infinity 1260 liquid chromatograph and diode array detector (Agilent, USA), with monitoring at 210 nm, were used for the geranylgeraniol determination. A LiChrospher 100RP-18 column (12,5 cm long, 4 mm d. i. and 5 µm particle diameter; Merck, Darmstadt) was used, and the mobile phase was composed of methanol: 20 mM ammonium acetate (90:10, v/v), in an isocratic system.

4.2.3. Molecular analysis

DNA was extracted from recently expanded young leaves according to Doyle and Doyle (1990). DNA was quantified by comparison with known concentrations of standard DNA (lambda DNA; Invitrogen, Carlsbad, CA, USA) in electrophoresis agarose gels (1%) stained with GelRed (Biotium, Fremont, CA, USA).

Sixteen SSR markers developed for *B. orellana* (BorA2, BorA3, BorA5, BorB1, BorB4, BorB5, BorB12, BorC5, BorD1, BorD2, BorF9, BorG4, BorG11, BorH3, BorH7, BorH10) (Dequigiovanni *et al.* 2014, Dequigiovanni *et al.* in press) were used in the present study. These markers were selected based on their polymorphism and compatibility for multiplexing. An M13 sequence tail was added to the 5' end of each forward primer following a labeling protocol (Schuelke, 2000). Microsatellite fragments were amplified using a MyCycler Thermal Cycler (Bio-Rad, Hercules, CA, USA) in a total reaction volume of 10 µL, containing 20 ng of genomic DNA template, 1 U *Taq* DNA polymerase (Fermentas, Vilnius, Lithuania), 1X polymerase chain reaction buffer (10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.08% Nonidet P40), 0.25 mM each dNTP, 1.5 mM $MgCl_2$, 2.5 pmol of forward and M13 label primers (FAM, HEX or NED dyes) and 5 pmol of reverse primers.

Polymerase chain reactions were carried out according to Schuelke (2000), consisting of 94 °C (5 min), then 30 cycles at 94 °C (30 s)/T_a °C (45 s) (T_a= annealing temperature)/ 72 °C (45 s), followed by 8 cycles at 94 °C (30 s)/53 °C (45 s)/72 °C (45 s), and a final extension at 72 °C for 10 min. Quality of amplification was checked by electrophoresis in agarose gels (1.5%) stained with GelRed (Biotium). Fragment separation and detection were performed on an ABI Prism 3130xl capillary sequencer using GeneScan 500 Rox-labelled size standard (Applied Biosystems, Foster City, CA, USA). SSR patterns were scored using the Genemapper v4.0 software (Applied Biosystems).

4.2.4. Statistical analysis

4.2.4.1 Phytochemical diversity analyses

The dispersion of phenotypic diversity among accessions was evaluated by Principal Components Analysis (PCA), carried out with the R Statistical Software - ade4 package (Dray and Dufour, 2007). Cluster analysis was performed using Euclidean distances, and the UPGMA (Unweighted pair group method with arithmetic mean) clustering method,

implemented in R Statistical Software - stats package (R Core Team, 2015) and formatted with FigTree (Rambaut and Drummond, 2010).

4.2.4.2 Genetic diversity and structure analyses

Genetic diversity parameters, including total number of alleles (*A*), allelic richness (*Ar*), observed (H_O) and expected (H_E) heterozygosities, number of private alleles (*Ap*) and inbreeding coefficient (F_{IS}), were estimated for each locus using diveRsity package from the R project (Keenan *et al.*, 2013). Genetic distances between individuals were estimated by Rogers' genetic distance (Rogers, 1972). The resulting distance matrix was used to construct a dendrogram with the Neighbour-joining algorithm (Saitou and Nei, 1987), with 1,000 bootstrap replicates, implemented in Population 1.2.32 (Langella, 2002). The final dendrogram was formatted with FigTree (Rambaut and Drummond, 2010). The dispersion of genetic diversity among accessions was evaluated by Principal Coordinate Analysis (PCoA), carried out with the R Statistical Software - ade4 package (Dray and Dufour, 2007). The apparent outcrossing rate ($\hat{t}_a$) was estimated by using the estimated inbreeding coefficient according to the equation $\hat{t}_a = (1-f)/(1+f)$ (Weir, 1996).

We also analyzed the population structure and detection of admixture using a Bayesian model based on the clustering method implemented in Structure 2.3.4 (Pritchard *et al.*, 2000). An admixture model with correlated allele frequencies without prior population information was used. A burn-in period of 250,000 was used, followed by 500,000 Markov Chain Monte Carlo (MCMC) permutations. Ten replicates (runs) were carried out for each possible value of K (from 1 to 10). Two different approaches were used to detect the most likely K value: the first was that proposed by Pritchard *et al.* (2000) and the second proposed by Evanno *et al.* (2005), using the web based Structure Harvester v.0.6.92 (Earl and VonHoldt 2012). Hierarchical distribution of genetic variation within and among groups of annatto accessions was evaluated using “locus-by-locus” AMOVA with GenAlEx version 6.5 (Peakall and Smouse 2012), with individuals clustered into groups according to the Structure analysis. Significance was assessed by conducting 10,000 permutations.

4.3. Results

4.3.1. Phytochemical compound characterization

Seed moisture content of samples was used to achieve uniformity of other parameters for correlation between different samples. Moisture content ranged from 5.78 to 14.32 g (100 g DM)$^{-1}$ (Table 4.1). Lipid concentrations ranged from 2.14 g to 7.11 g (100 g DM)$^{-1}$. Total carotenoids, expressed as bixin, showed concentrations (dry basis) ranging from a minimum of 2.00 g to a maximum of 7.31 g (100 g DM)$^{-1}$.

Analyses of tocotrienols identified the predominance of γ-tocotrienol and δ-tocotrienol, with δ-tocotrienol representing approximately 90% of the observed isoforms. The α and β-tocotrienols were not present or were below the analytical sensitivity limit of the method used [0.01 g (100 g DM)$^{-1}$]. Total tocotrienols concentrations (dry basis) ranged from a minimum of 0.25 g to a maximum of 1.05 g (100 g DM)$^{-1}$ (Table 4.1). Analyses of geranylgeraniol showed results ranging from 0.49 g to 2.61 g (100 g DM)$^{-1}$.

The cluster analysis conducted with phytochemical compounds classified the accessions into seven groups (Figure 4.2). The accessions from São Paulo, Rio de Janeiro and Mato Grosso all clustered in the first three groups, which together showed low values for all traits. The accessions from Rondônia clustered in the other four groups, with the exception of accessions 63 (group 1) and 54 (group 2). These groups also contained one accession from Minas Gerais and eight of unknown origin. These four groups showed higher values than the other three groups for all traits. Moreover, within these four groups there are accessions individually responsible for the highest values in all characteristics, such as accessions 62 (Bixin: 7.11), 56 (Geranylgeraniol: 2.61), 47 and 23 (Tocotrienol: 1.05).

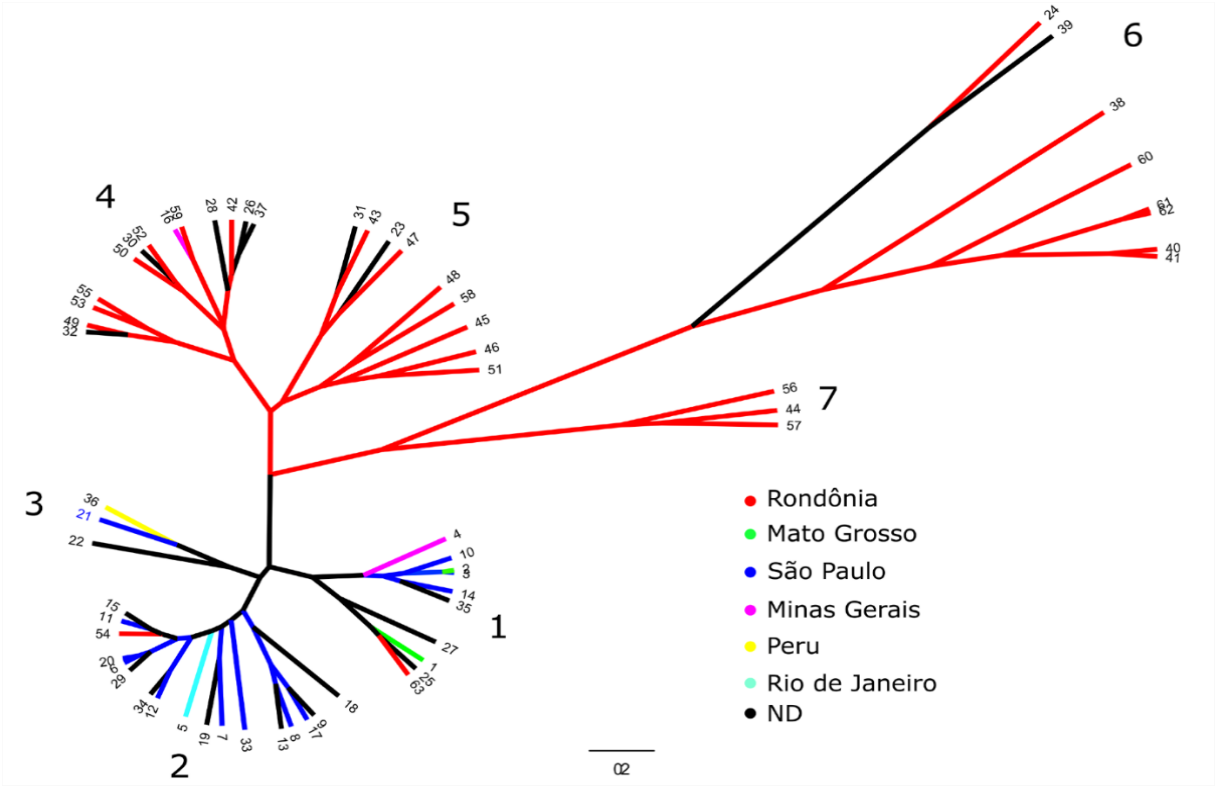

Figure 4.2 Cluster analysis using Euclidean distances and the UPGMA algorithm based on four phytochemical compounds analyzed in 63 *Bixa orellana* accessions from the Brazilian States of Rondônia, Mato Grosso, São Paulo, Minas Gerais, Rio de Janeiro, from Peru and 20 with unknown origins (ND)

The PCA explained nearly 83% of the variation in the data set (Figure 4.3). PCA1 can be considered an axis of phytochemical constituent concentration, with higher values to the left, attributed mostly to Rondônia accessions, while PCA2 distinguishes between annattos with more geranylgeraniol and lipids above the axis, and those with more tocotrienols and bixin below the axis.

Significant linear correlations were observed ($P < 0.05$) between the geranylgeraniol concentration and lipids (r = 0.63), bixin and lipids (r = 0.51), tocotrienols and lipids (r = 0.41) and between bixin and tocotrienols (r = 0.56). There were no correlations between geranylgeraniol and bixin, nor between geranylgeraniol and tocotrienols.

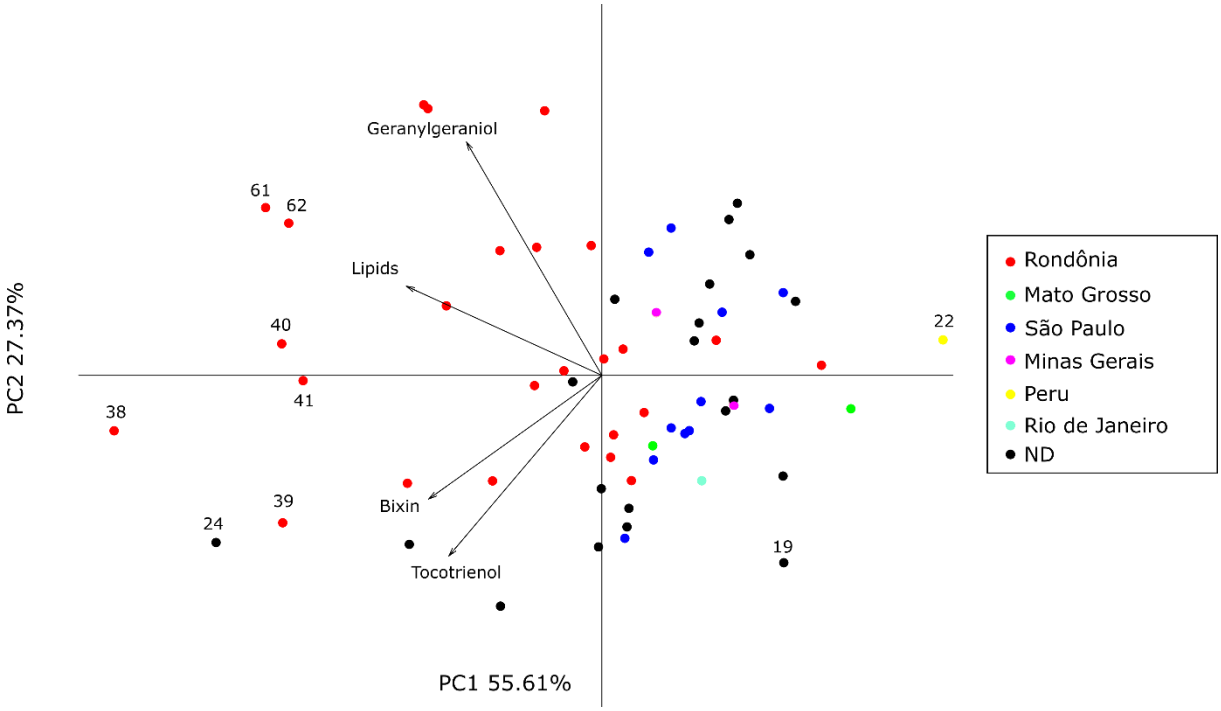

Figure 4.3 Principal component analysis biplot of the 63 *Bixa orellana* accessions based on bixin, tocotrienol, geranylgeraniol and lipids analysis

4.3.2. Molecular characterization

The characterization of 63 accessions with 16 microsatellite loci identified a total of 73 alleles, varying from two to ten alleles per locus, with an average of 4.56 alleles per locus (Table 4.2). We found higher H_E then H_O values for all loci, indicating heterozygote deficits. As a result, F_{IS} was high for most loci and 14 loci were not in Hardy-Weinberg Equilibrium. This result is expected for genebanks where populations are not sampled, but individuals from diverse populations are grouped in a common area, disagreeing with the Hardy-Weinberg premises that population should be panmictic and with infinite size. Polymorphism Information Content (PIC) varied widely, with an average of 0.43.

In the PCoA analysis, the first two principal coordinates explained 48.3% of total variation and showed that accessions from Southeast and Central-West Brazil tend to form a separate group from the North (Rondônia) accessions, which were more dispersed and clustered mainly towards the right side of the first principal coordinate (Figure 4.4). Cluster analysis based on Rogers' distance and the Neighbour-Joining algorithm (Figure 4.5) presented two major clusters, revealing the same pattern observed in the PCoA. All accessions from Rondônia State, except two, were grouped close to each other, while accessions from Southeast Brazil formed a distinct group. The two Central-West (Mato Grosso) accessions

were distributed in both major clusters. Bayesian analysis performed in Structure also confirmed the results obtained with PCoA and the NJ dendrogram. Two genetic clusters were obtained (ΔK = 776.35), according to the Evanno method (Figure 4.6), showing that most of the accessions in the red group belong to the Southeast region, while those of the yellow group, with three exceptions, and excluding Peru and the non-determined origin accessions, were from northern Rondônia (Figure 4.7).

Table 4.2 Genetic parameters for the microsatellite analysis of *Bixa orellana* accessions in the Instituto Agronômico germplasm bank with 16 SSR loci: *A*, number of alleles; H_O, observed heterozigosity; H_E, expected heterozigosity; F_{IS}, inbreeding coefficient; PIC, polymorphism information content

SSR Loci	*A*	H_O	H_E	F_{IS}	PIC
BorA2	3	0.16	0.61	0.73	0.53
BorA3	3	0.12	0.35	0.65	0.30
BorA5	3	0.08	0.18	0.53	0.16
BorB12	4	0.22	0.52	0.56	0.46
BorF9	8	0.28	0.76	0.62	0.72
BorG11	7	0.29	0.76	0.62	0.72
BorB1	6	0.24	0.63	0.62	0.58
BorB4	3	0.17	0.51	0.66	0.39
BorB10	10	0.31	0.83	0.63	0.80
BorC12	3	0.10	0.37	0.74	0.31
BorD1	3	0.27	0.37	0.28	0.31
BorD2	2	0.00	0.32	1.00	0.26
BorG4	5	0.16	0.32	0.50	0.29
BorH3	2	0.03	0.03	0.01	0.03
BorH7	5	0.13	0.45	0.71	0.43
BorH10	6	0.24	0.57	0.58	0.53
Overall Mean	4.56	0.17	0.47	0.63	0.43

Results of AMOVA, using individuals clustered into groups according to the Structure analysis, showed that most of the genetic variation resided within groups (89%), while 11% of the variation resided between groups. Divergence was moderate between clusters (F_{ST} = 0.112, p = 0.000). The results of the PCoA, NJ tree and Structure showed signs of admixture (Figures 4.4, 4.5 and 4.7).

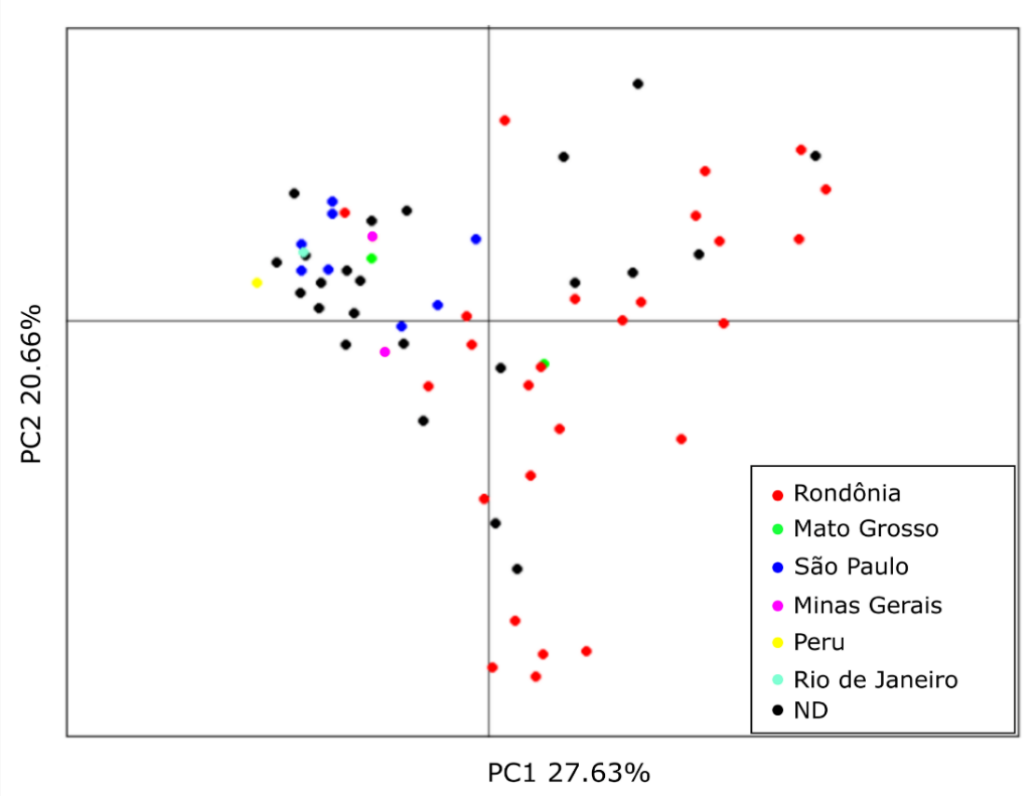

Figure 4.4 Principal coordinate analysis of the dispersion of the 63 *Bixa orellana* accessions evaluated with 16 SSR markers

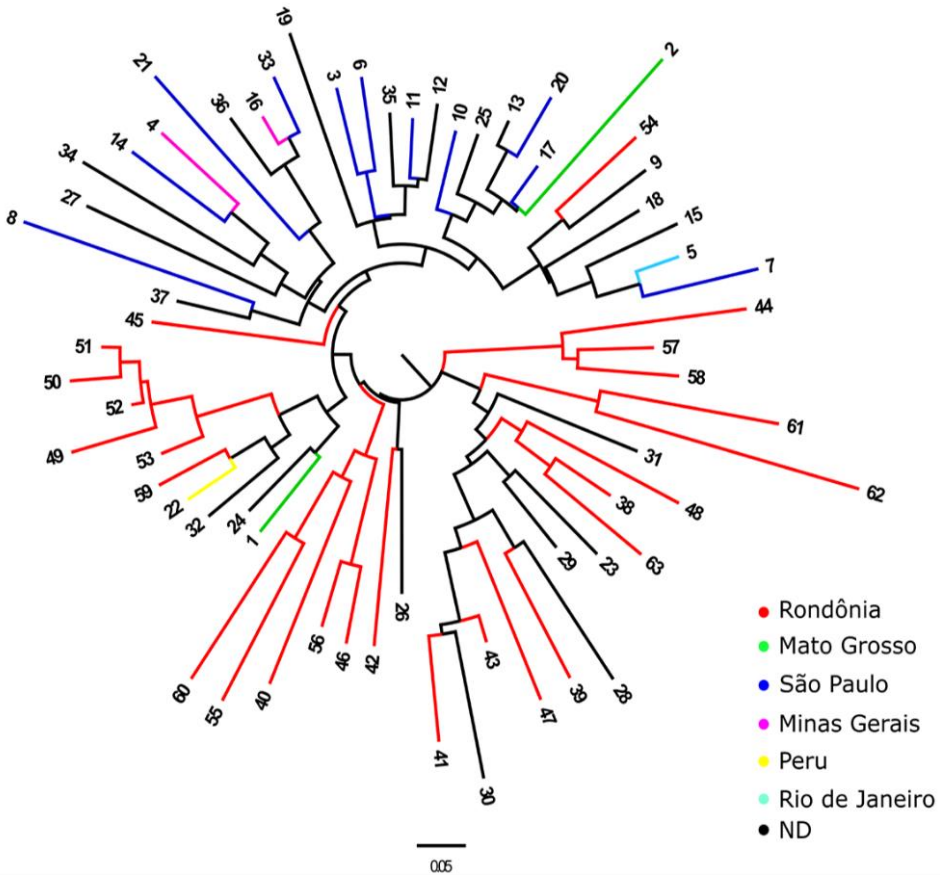

Figure 4.5 Neighbour-Joining tree of 63 *Bixa orellana* accessions (colored by origin), constructed with Rogers' distance based on 16 SSR loci

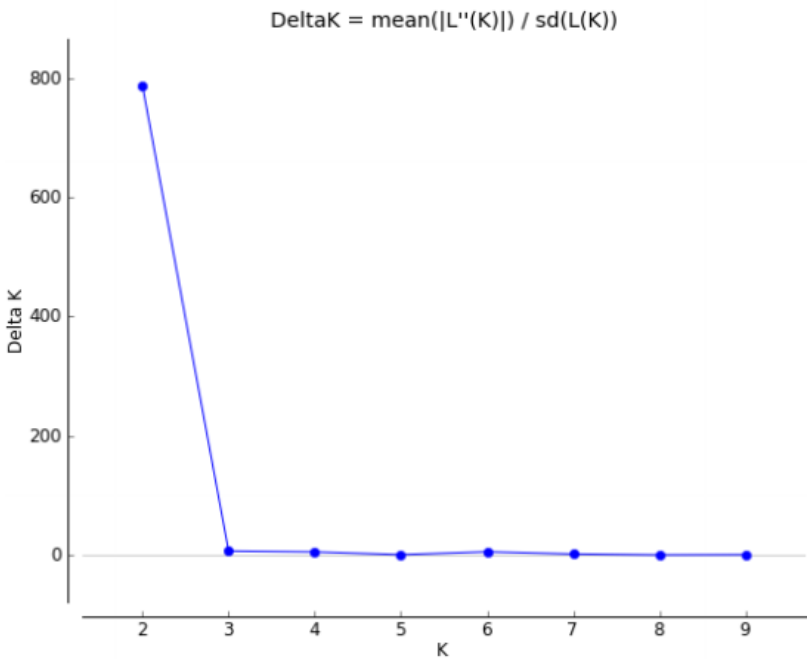

Figure 4.6 Evanno *et al.* (2005) plot detecting the number of *K* groups that best fit the data

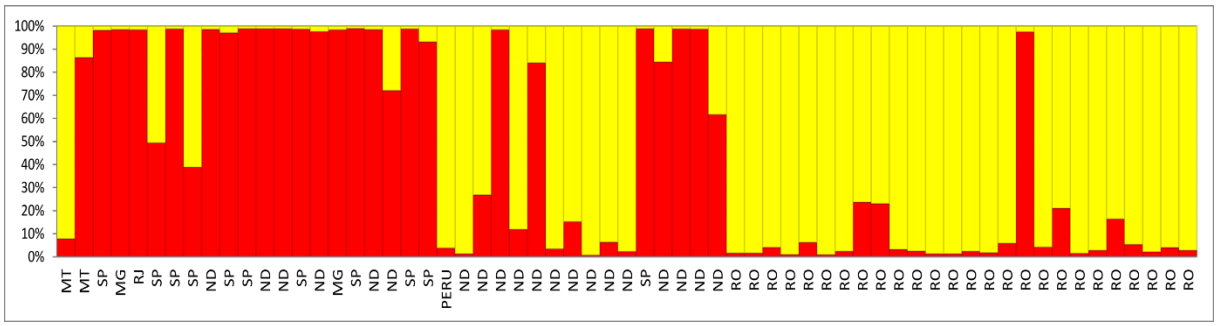

Figure 4.7 The estimated proportion of membership in the corresponding clusters (K = 2), calculated using Structure, for 63 accessions of *Bixa orellana* evaluated with 16 SSR loci

4.4. Discussion

4.4.1. Phytochemical characterization

The annatto accessions analyzed in this study showed moisture content values ranging from 5.78 g to 14.32 g (100 g DM)$^{-1}$. Stringheta and Silva (2008) established 14 g (100 g)$^{-1}$ as the maximum moisture content of annatto seeds, above which there was a risk of mold growth. Only one accession (sample 45) showed a value above 14 g (100 g DM)$^{-1}$, which shows that the analyzed accessions present adequate moisture contents. As for the lipid contents, the analyzed accessions showed lipid values in agreement with Frega *et al.* (1998) and Rao *et al.* (2015), reporting an average of 5.3 g (100g DM)$^{-1}$ and 6.3 g (100g DM)$^{-1}$, respectively.

Currently, the main criterion of seed quality used for the improvement of this crop is the concentration of bixin, which can be as high as 5.0 g (100 g DM)$^{-1}$ (Vilar *et al.*, 2014). According to Franco *et al.* (2008), levels higher than 2.5 g (100 g DM)$^{-1}$ are usually required for export. Annatto accessions maintained by this germplasm bank showed bixin concentrations ranging from 2.0 to 7.1 g (100 g DM)$^{-1}$, where most of the accessions (93.6%) produced values higher than 2.5 g (100 g DM)$^{-1}$, and therefore have export potential. This wide variation in bixin content in different varieties of annatto was also observed by Matos *et al.* (1992), Carvalho *et al.* (2010) and Mantovani *et al.* (2013).

Total tocotrienols concentration of these annatto accessions, ranging from 0.25 g to 1.05 g (100g DM)$^{-1}$, were much higher than the mean concentration of 0.14 g (100 g DM)$^{-1}$ reported by Frega *et al.* (1998). This variability shows the importance and viability of conducting studies to select varieties with high concentrations of tocotrienols. Geranylgeraniol is the major oily constituent of annatto seeds, representing 1% of dry seeds (Vilar *et al.*, 2014). Analyses of geranylgeraniol in annatto seeds in this study showed values ranging from 0.49 g to 2.61 g (100 g DM)$^{-1}$. Most of the accessions showed values well above those obtained by Smith and Wallin (2006) [average of 0.75 g (100 g DM)$^{-1}$]. As in the case of tocotrienols, the wide variation of the concentration of geranylgeraniol present in the annatto accessions in this germplasm collection indicates the need to conduct studies selecting varieties with a high concentration of this component.

This study showed that the Northern accessions generally have high concentrations of the phytochemical compounds, which leads us to speculate that this finding could be related to the Amazonian origin and domestication of annatto (Sandy-Cuen and Becerra, 2003; Clement *et al.*, 2010; Moreira *et al.*, 2015), yet to be confirmed. Accession 63, which morphologically appears to be a wild accession, presenting small fruits, fewer seeds per fruit and seeds only partially covered with aril, presented low values for all phytochemical characteristics, consistent with what would be expected for wild accessions. Due to the fact that wild accessions have not suffered human selection for traits of human interest, it is expected that it should present lower values compared to the domesticated materials (Moreira *et al.*, 2015).

4.4.2. Molecular characterization

Heterozygosity is a measure of the genetic variability that estimates how much variation exists in a population and how this variation is distributed depending on the alleles present at a given locus (Frankham *et al.*, 2004). The expected heterozygosity values, with a mean of 0.47, exceeded those obtained for the observed heterozygosity, averaging 0.17, for all loci in this study. When plants from different populations are grouped and analyzed in a single group, as with accessions from germplasm banks, the expected heterozygosity is usually greater than the observed heterozygosity, which is known as the Wahlund effect (Hartl and Clark, 1998). High values of F_{IS} observed in all loci analyzed allowed the interpretation that the genotypes studied are also subjected to inbreeding. From these values, we calculated the apparent outcrossing rate ($\hat{t}_a$), yielding a value of 0.22. From this value, it is possible to interpret that the apparent selfing rate of the plants collected for representation in the germplasm bank is 0.78. The existence of crosses between closely related individuals generates inbreeding and, therefore, increases homozygosity (Ritland, 2002). These results are in agreement with those obtained by Rivera-Madrid *et al.* (2006), conducting controlled pollinations in annatto accessions in an experimental field in Mexico, suggesting that annatto can tolerate both types of pollination, with recorded cross-pollination rates of 57% and self-fertilization of 31.4%. Vilares *et al.* (1992) also concluded that there was natural selfing in annatto. It is worth mentioning that apparent outcrossing rate and apparent selfing rate are estimates, and do not replace progeny tests to determine the mating system, and therefore need further studies to confirm these data.

In all the molecular analyses, such as the cluster and PCoA analyses, the Rondônia accessions, with only a few exceptions, were classified into groups separated from the Southeastern accessions of Brazil, similar to the phytochemical characterization. They were also clearly classified in one group separated from the other accessions from Brazil in the Bayesian analysis, which classified the accessions into two groups, showing genetic differentiation among the accessions in Brazil, with a genetic structuring separating the Northern from the Southeastern accessions. These results are in agreement with those obtained by Carvalho *et al.* (2005), which presented a clear differentiation between Northern accessions from those obtained in other regions of the country. Annatto is native to the Americas, and most probably the Amazon region (Sandy-Cuen and Becerra, 2003; Clement *et al.*, 2010), which might explain the higher levels observed for the phytochemical traits and genetic diversity values from the Northern (Rondônia) accessions.

In conclusion, our data revealed high variation of phytochemical compounds in the 63 accessions from the annatto Germplasm Bank of IAC, corroborated by high levels of genetic diversity revealed by 16 SSR markers. Interestingly, Northern accessions concentrate higher concentrations of the phytochemical compounds and higher levels of genetic diversity. Some accessions presented bixin levels well above average. Considering that this trait is a key feature in this crop, these materials are very promising to be used in genetic improvement programs. Additional studies with local varieties, as well as wild populations of annatto collected in several regions in Brazil, are underway by our team and they should contribute to a better understanding of the distribution of genetic diversity in this country, and further information concerning the center of domestication of *Bixa orellana*.

REFERENCES

Akshatha, V., Giridhar, P. and Ravishankar, G.A. (2011) Morphological diversity in *Bixa orellana* L. and variations in annatto pigment yield. *J. Hortic. Sci. Biotechnol.* **86**, 319–324.

Albuquerque, C.L.C. and Meireles, M.A.A. (2012) Defatting of annatto seeds using supercritical carbon dioxide as a pretreatment for the production of bixin: Experimental, modeling and economic evaluation of the process. *J. Supercrit. Fluids* **66**, 86–94.

Alves, R.W., Ulson De Souza, A.A., Ulson De Souza, S.M.D.A.G. and Jauregi, P. (2006) Recovery of norbixin from a raw extraction solution of annatto pigments using colloidal gas aphrons (CGAs). *Sep. Purif. Technol.* **48**, 208–213.

Arce, J. (1999) El achiote *Bixa orellana* L. cultivo promisorio para el trópico. In Earth., p. 149.

Auttachoat, W., Germolec, D.R., Smith, M.J., White, K.L. and Guo, T.L. (2011) Contact sensitizing potential of annatto extract and its two primary color components, cis-bixin and norbixin, in female BALB/c mice. *Food Chem. Toxicol.* **49**, 2638–2644.

Baer, D.. (1976) *Systematics of the genus Bixa and geography of the cultivated annatto*. University of California - Los Angeles.

Carvalho, J.F.R.P. de, Robinson, I.P. and Alfenas, A.C. (2005) Isozymic variability in a Brazilian collection of annatto (*Bixa orellana* L.). *Pesqui. Agropecuária Bras.* **40**, 653–660.

Carvalho, P.R.N., Carvalho, C.R.L. and Mantovani, D.M.B. (1991) Estudo da composição de sementes, cachopas, folhas e galhos do urucueiro. In Seminário Internacional de Corantes Naturais Para Alimento., p. 317.

Carvalho, P.R.N., Silva, M.G. da, Fabri, E.G., Tavares, P.E.R., Martins, A.L.M. and Spatti, L.R. (2010) Concentração de bixina e lipidios em sementes de urucum da coleção do Instituto Agronômico (IAC). *Bragantia* **69**, 519–524.

Carvalho, P.R.N., Silva, M.G. da and Moreira, C.G.C. (1993) Avaliação dos métodos espectrofotométricos de análise de sementes de urucum (*Bixa orellana* L.). *Colet. Inst. Tecnol. Aliment.* **23**, 181–8.

Clement, C.R., Cristo-Araújo, M. de, D'Eeckenbrugge, G.C., Pereira, A.A. and Picanço-Rodrigues, D. (2010) Origin and domestication of native Amazonian crops. *Diversity* **2**, 72–106.

Craveiro, A.R., Oliveira, C.L.A. and Araujo, F.W.L. (1989) The presence of geranylgeraniol in *Bixa orelana*, Linn. *Quim. Nova* **12**, 297–298.

Dequigiovanni, G., Ramos, S.L.F., Zucchi, M.I., Bajay, M.M., Pinheiro, J.B., Fabri, E.G., Bressan, E.A. and Veasey, E.A. (2014) Isolation and characterization of microsatellite loci for *Bixa orellana*, an important source of natural dyes. *Genet. Mol. Res.* **13**.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of Plant DNA from fresh tissue. *Focus (Madison).* **12**, 13–15.

Dray, S. and Dufour, A.-B. (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. *J. Stat. Softw.* **22**, 1–20.

Earl, D.A. and vonHoldt, B.M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361.

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**, 2611–20.

Fabri, E.G. (2015) Demanda por corantes naturais aquece mercado brasileiro de urucum. *Soc. Nac. Agric.*

Franco, C.F. de O., Fabri, E.G., Barreiro Neto, M., Manfiolli, M.H., Harder, M.N.C. and Rucker, N.C. de A. (2008) Urucum: sistemas de produção para o Brasil, João Pessoa : EMEPA-PB.

Frankham, R., Ballou, J.D. (2004) A primer of conservation genetics, Cambridge University Press.

Frega, N., Mozzon, M. and Bocci, F. (1998) Identification and estimation of tocotrienols in the annatto lipid fraction by gas chromatography-mass spectrometry. *J. Am. Oil Chem. Soc.* **75**, 1723–1727.

Giorgi, A., Marinis, P. De, Granelli, G., Chiesa, L.M. and Panseri, S. (2013) Secondary metabolite profile, antioxidant capacity, and mosquito repellent activity of *Bixa orellana* from Brazilian Amazon region. *J. Chem.*

Giuliano, G., Rosati, C. and Bramley, P.M. (2003) To dye or not to dye: biochemistry of annatto unveiled. *Trends Biotechnol.* **21**, 513–6.

Hartl, D. and Clark, A. (1998) Principles of Population Genetics, {Sinauer Associates}.

Horwitz, W. (2005) Official Methods of Analysis of the Association of Official Analytical Chemists 18th ed. HORWITZ, W., ed,.

Jondiko, I.J.O. and Pattenden, G. (1989) Terpenoids and an apocarotenoid from seeds of *Bixa orellana*. *Phytochemistry* **28**, 3159–3162.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R. and Dhawan, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177**, 309–334.

Keenan, K., McGinnity, P., Cross, T.F., Crozier, W.W. and Prodöhl, P.A. (2013) diveRsity : An R package for the estimation and exploration of population genetics parameters and their associated errors O'Hara, R.B., ed. *Methods Ecol. Evol.* **4**, 782–788.

Langella, O. (2002) Populations, 1.2.32. Population genetic software.

Leal, F. and Clavijo, C.M. (2010) Acerca de la história, taxonomia, botánica y usos de Bixa orellana L. *Rev. Unell. Cienc. Tec* **1**, 78–86.

Leal, F. and Clavijo, C.M. de (2012) Annatto: Botany and Horticulture. In Horticultural Reviews., pp. 389–419. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Lopes, M.V., Desoti, V.C., Caleare, A.D.O., Ueda-Nakamura, T., Silva, S.O. and Nakamura, C.V. (2012) Mitochondria superoxide anion production contributes to geranylgeraniol-induced death in leishmania amazonensis. *Evidence-based Complement. Altern. Med.* 298320.

Mantovani, N.C., Grando, M.F., Xavier, A. and Otoni, W.C. (2013) Avaliação de genótipos de urucum (*Bixa orellana* L.) por meio da caracterização morfológica de frutos, produtividade de sementes e teor de bixina. *Ciência Florest.* **23**, 344–362.

Matos, F.J.A., Alencar, J.W., Craveiro, A.A. and Machado, M.I.L. (1992) Acidos graxos de algumas oleaginosas tropicais em ocorrência no nordeste brasileiro. *Quim. Nova* **15**, 181–195.

Monzote, L., Montalvo, A.M., Almanonni, S., Scull, R., Miranda, M. and Abreu, J. (2006) Activity of the essential oil from *Chenopodium ambrosioides* grown in Cuba against *Leishmania amazonensis*. *Chemotherapy* **52**, 130–6.

Moreira, P.A., Lins, J., Dequigiovanni, G., Veasey, E.A. and Clement, C.R. (2015) The Domestication of Annatto (*Bixa orellana*) from *Bixa urucurana* in Amazonia. *Econ. Bot.* **69**, 127–135.

Nisar, N., Li, L., Lu, S., *et al.* (2015) Carotenoid metabolism in plants. *Mol. Plant* **8**, 68–82.

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.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

R Core Team (2015) R: A language and environment for statistical computing.

Rambaut, A. and Drummond, A. (2010) FigTree v1.3.1.

Rao, P., Narsing Rao, G., Jyothirmayi, T., *et al.* (2015) Characterisation of Seed Lipids from *Bixa orellana* and *Trachyspermum copticum*. *JAOCS, J. Am. Oil Chem. Soc.* **92**, 1483–1490.

Ritland, K. (2002) Extensions of models for the estimation of mating systems using n independent loci. *Heredity (Edinb).* **88**, 221–8.

Rivera-Madrid, R., Escobedo-GM, R.M.M., Balam-Galera, E., Vera-Ku, M. and Harries, H. (2006) Preliminary studies toward genetic improvement of annatto (*Bixa orellana* L.). *Sci. Hortic. (Amsterdam).* **109**, 165–172.

Rodrigues, J. (1995) *Análise de isoenzimas em progênies de meio-irmãos de urucum (Bixa orellana L.)*. Federal University of Viçosa - UFV.

Rogers, J. (1972) Measures of genetic similarity and genetic distances. In Studies in Genetics VII., pp. 145–153. Austin, TX: University of Texas Publication.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25.

Sandy-Cuen, P.M. and Becerra, R. (2003) Manejo campesino de recursos naturales. El achiote. *BioDiversitas* **7**, 7–11.

Schuelke, M. (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18**, 233–234.

Sen, C.K., Khanna, S. and Roy, S. (2007) *Tocotrienols in health and disease: The other half of the natural vitamin E family,*.

Smith, J. and Wallin, H. (2006) *Annatto extracts. Chemical and technical assessment,*.

Stringheta, C.P. and Silva, P.I. (2008) Pigmentos de urucum. Extração, Reações Químicas, Uso e Aplicações 1st ed., Viçosa, MG: Suprema.

Tan, B. and Foley, J. (2000) Tocotrienols and geranylgeraniol from *Bixa orellana* byproducts. *WO Pat. 2,000,071,531*.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. (2016) Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **39**, 312–328.

Vilar, D. de A.A., Vilar, M.S. de A.S. de A., Lima e Moura, T.F.A. de, *et al.* (2014) Traditional uses, chemical constituents, and biological activities of *Bixa orellana* L.: a review. *Sci. World J.* **857292**, 1–11.

Vilares, A.S., São José, A.R., Rebouças, T.N.H. and Souza, I.V.B. (1992) Estudo da biologia floral de urucuzeiro (*Bixa orellana* L.). *Rev. Bras. Corantes Nat.* **1**, 101–105.

Weir, B.S. (1996) Genetic data analysis II: methods for discrete population genetic data, Sinauer Associates.

Zahn, T.J., Eilers, M., Guo, Z., Ksebati, M.B., Simon, M., Scholten, J.D., Smith, S.O. and Gibbs, R.A. (2000) Evaluation of isoprenoid conformation in solution and in the active site of protein-farnesyl transferase using carbon-13 labeling in conjunction with solution- and solid-state NMR. *J. Am. Chem. Soc.* **122**, 7153.

5. GENETIC DIVERSITY AND DISTRIBUTION OF ANNATTO (*BIXA ORELLANA* L.) FROM AMAZONIA AND SOUTHEAST OF BRAZIL

Abstract

Annatto (*Bixa orellana* L.) is a small evergreen tropical American tree native to the northern South America. Historical documents report its extensive distribution and cultivation in the American tropics before European conquest, and its subsequent dispersal to the rest of the tropical world. Brazil has the greatest morphological and genetic diversity of the species, and is also the major producer of annatto. Considering the lack of information on the genetic diversity and structure of *B. orellana*, this study aimed to characterize the amount and distribution of genetic diversity among cultivated samples of annatto from Brazilian Amazonia, collected from homegardens of riverside communities along the Rivers Madeira, Negro, Branco, Solimões and Amazon, and from farmer´s fields along highways, in the States of Amazonas, Rondônia, Roraima and Pará, and Southeastern Brazil, specifically the State of São Paulo, using 16 highly polymorphic microsatellite markers. The samples collected presented ample morphological variation in fruits color, shape and size. Moderate levels of genetic diversity were observed among the annatto groups, highlighted by the mean number of alleles per locus ($\bar{A}$ = 5.3) and the mean observed (H_O = 0.141) and expected heterozygosities (H_E = 0.462). Lower H_O values compared to H_E values were observed for all geographic groups analysed. Moderate to high levels of admixture between geographic groups were observed. However, neighbor-joining dendrograms, principal coordinate and Bayesian analysis showed a tendency to group annatto samples based on their geographic origin. A significant positive correlation between genetic and geographic distances according to Mantel's test indicates that annatto plants may be experiencing isolation by distance. We suggest that efforts to preserve the existing variability of annatto in the riverside communities along the main rivers of Brazilian Amazonia must be made, as well as in the State of Pará, using strategies of *in situ* conservation. These areas showed higher genetic diversity besides presenting a high number of private alleles.

Keywords: Amazonia; Microsatellite; Genetic structure; Conservation; SSR

5.1. Introduction

Annatto (*Bixa orellana* L.) is a small evergreen tropical American tree (Arce, 1999) native to the northern South America (Sandy-Cuen and Becerra, 2003; Clement *et al.*, 2010). Historical documents report its extensive distribution and cultivation in the American tropics before European conquest, and its subsequent dispersal to the rest of the tropical world (Leal and Clavijo, 2010; Akshatha *et al.*, 2011). The botanical name of this species is entirely historical in its significance. The genus name comes from the Taino word, "bixa", which was pronounced by the Native Americans as "bisha", while the name "Orellana" comes from the Spanish conquer, Francisco de Orellana, who is credited with discovering the Amazon River in 1541 (Morton, 1960). The Taino were the first people that Christopher Columbus met in

the Caribbean in 1492. The common name in Portuguese, “Urucum”, comes from Tupi-Guarani languages and means red (Baleé, 2000).

The main commercial producers of annatto are countries in South America (about 60 % of the world's annatto production), followed by Africa (27 %) and Asia (12 %) (Giuliano *et al.*, 2003). Brazil hosts the greatest morphological and genetic diversity of the species, and is also the major producer of annatto (Akshatha *et al.*, 2011). Annatto is very common in homegardens in the North and Northeast of Brazil, but is also grown in other regions. The states of Rondônia, São Paulo and Pará are the largest producers in the country (Franco *et al.*, 2008; Akshatha *et al.*, 2011).

Annatto is commercially valuable because it is the only source of the natural apocarotenoid “bixin”, the principal coloring constituent of annatto pigment. Bixin is produced in the arils of the seeds and accounts for more than 80 % of the total carotenoid content (Rivera-Madrid *et al.*, 2006). This orange-red pigment is widely used in the food and cosmetics industries instead of synthetic pigments (Giuliano *et al.*, 2003; Nisar *et al.*, 2015). The demand for annatto increased with the prohibition of the use of the synthetic dye *Tartrazine* in several countries (Andres, 1980). In addition, annatto has been recognized as a highly desirable colorant for cheese, butter, margarine and other dairy products, affecting neither the flavor nor the aroma of the product (Morton, 1960; Akshatha *et al.*, 2011). It ranks second in economic importance worldwide among all natural colorants and also contains other important substances for human health, such as geranylgeraniol, tocotrienols and other carotenoids presenting antimicrobial, antioxidant and antiviral properties (Cáceres *et al.*, 1995; Coelho *et al.*, 2003; Yolmeh *et al.*, 2014). Annatto has also shown potential in reducing levels of glucose and cholesterol in blood (Russell *et al.*, 2005). The residue from the bixin extraction process is a useful additive to poultry feed and can replace 30 % of the maize in animal feed (Akshatha *et al.*, 2011).

Annatto dye is also of great importance in the history of Brazil. Since ancient times, Native Americans in the Neotropics made extensive use of the colorful arils for body painting and dyeing of clothes, which immediately caught the attention of European explorers (Morton, 1960). Native South Americans believe that annatto can ward off evil spirits, and also as a protection from insects and from the ultra-violet rays of the sun (Morton, 1960; Plotkin, 1993). Annatto extracts from leaves, roots and seeds have traditionally been used for medicinal purposes, such as the treatment of wounds, diarrhea and asthma. The Mayas also used annatto for painting pottery (Morton, 1960).

There is still little information concerning the genetics of annatto. *B. orellana* is a cross-pollinated species, but can tolerate selfing (Rivera-Madrid *et al.*, 2006; Valdez-Ojeda *et al.*, 2010; Akshatha *et al.*, 2011; Lombello and Pinto-Maglio, 2014). The species has 2n = 14 chromosomes and presents one of the lowest DNA amount estimates (4C = 0.78 pg) for angiosperms (Ohri *et al.*, 2004; Lombello and Pinto-Maglio, 2014).

Despite its growing economic and historically importance, the conservation and manipulation of annatto genetic resources represent major challenges. Characterizing the genetic structure and diversity of cultivated populations of a crop is essential for the efficient use of its genetic resources (Clement *et al.*, 2010). Morpho-agronomic descriptors and molecular markers can be used to characterize this diversity. Microsatellite markers have been widely used to study the genetic diversity and genetic structure of plant populations (Kalia *et al.*, 2011; Vieira *et al.*, 2016). A total of 57 microsatellite loci have been developed for *B. orellana* by Dequigiovanni *et al.* (2014) and Dequigiovanni *et al.* (in press; Chapter 2), and 16 loci were used to evaluate accessions from a Brazilian annatto germplasm collection (Dequigiovanni *et al.*, in press; Chapter 4). Microsatellite markers or simple sequence repeats (SSR) are hypervariable and co-dominant, present highly reproducible results, and are widely distributed in the nuclear genome (Kalia *et al.*, 2011; Vieira *et al.*, 2016).

Considering the lack of information on the genetic diversity and structure of *B. orellana*, this study aimed to characterize the amount and distribution of genetic diversity among cultivated samples of annatto (*B. orellana*) from Brazilian Amazonia and Southeastern Brazil, specifically the State of São Paulo, using highly polymorphic microsatellite markers.

5.2. Material and methods

5.2.1. Plant material and DNA isolation

A total of 512 samples of annatto (*Bixa orellana*) were used in this study, collected from 37 municipalities in Brazil (Figure 5.1). Based on our field investigations from 2009 to 2016, cultivated annatto varieties were collected from homegardens of riverside communities in Amazonia along the upper Madeira River, middle Negro River, Solimões River and Amazon River in the states of Amazonas and Pará, and the Branco River in the state of Roraima. Due to the small number of samples and geographic proximity, the samples from the Branco and Negro Rivers were grouped into a single group. We also sampled in farmer´s fields along highways in the states of Rondônia, eastern Pará, and São Paulo. Samples from

southeastern Brazil and the state of Rondônia maintained by the Germplasm Bank of the Agronomic Institute of Campinas (IAC) were also included in the analysis.

From each sample, a leaf was collected and stored in plastic bags containing silica gel or in CTAB gels (3 % Cetyl Trimethyl Ammonium Bromide and 35 % NaCl). Geographical coordinates of each collection site were registered using a Global Positioning System (GPS). Total genomic DNA was extracted following Doyle and Doyle (1990) with minor modifications (CTAB 3 %). DNA concentration was determined by electrophoresis in 1 % agarose gels stained with GelRed (Biotium) and comparison with 25, 50 and 100 ng/µL of lambda phage DNA (Invitrogen). Uniform DNA concentrations of 5 ng/µL were then used for Polymerase Chain Reactions.

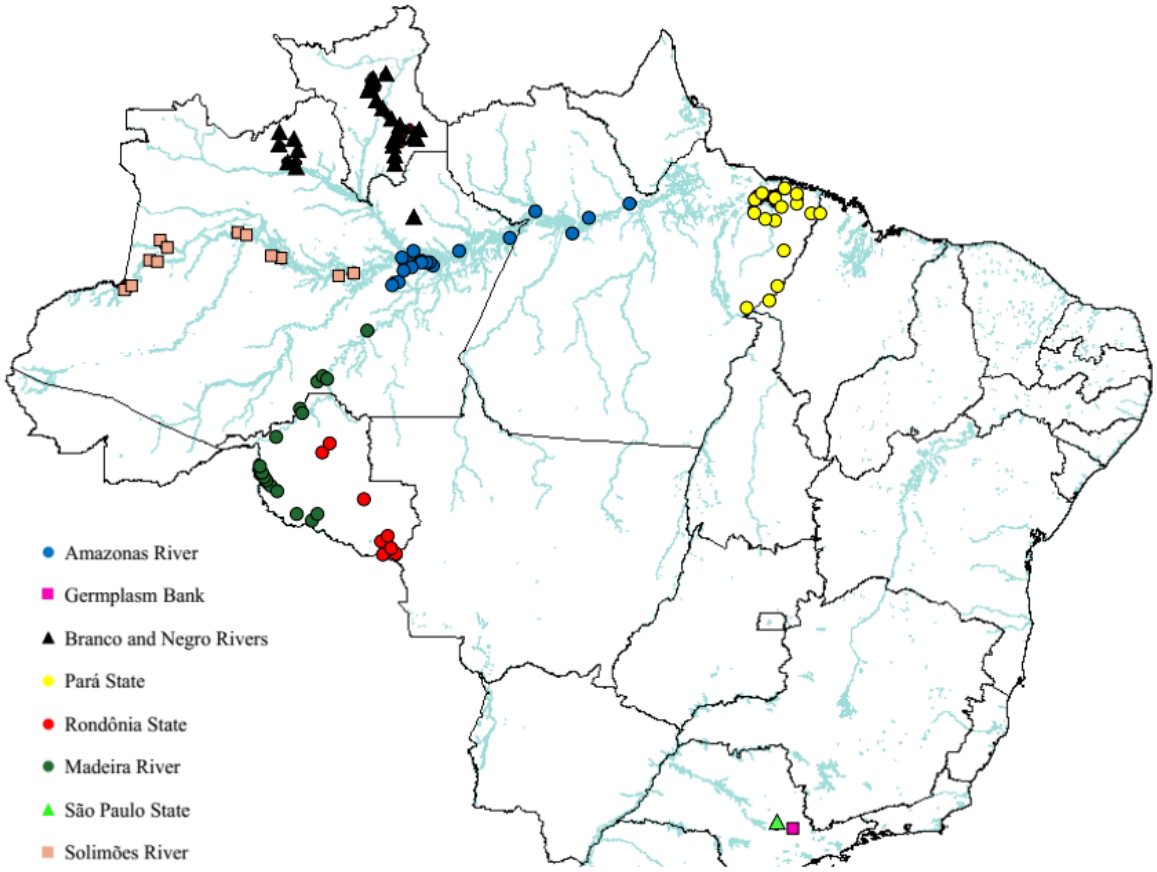

Figure 5.1. Geographic distribution of collection localities of cultivated *Bixa orellana* samples in Brazilian Amazônia and the State of São Paulo

5.2.2. PCR amplification and genotyping of SSRs

Sixteen fluorescently-labeled SSR loci (Dequigiovanni *et al.*, 2014; Dequigiovanni *et al.*, in press; Table 5.2) were used in this study. For each marker, forward primers were fluorescently labelled with NED, FAM or HEX. Polymerase chain reactions were carried out

in a MyCycler Thermal Cycler (Bio-Rad, Hercules, CA, USA) and reaction mixtures contained 20 ng of genomic DNA template, 1 U *Taq* DNA polymerase (Fermentas, Vilnius, Lithuania), 1X polymerase chain reaction buffer (10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.08 % Nonidet P40), 0.25 mM each dNTP, 1.5 mM $MgCl_2$, 2.5 pmol of forward and M13 label primers (FAM, HEX or NED dyes) and 5 pmol of reverse primers in a total volume of 10 µL. The amplification reactions for each locus were conducted according to Schuelke (2000) as follows: an initial denaturing step at 94 °C for 5 min, followed by 30 cycles at 94 °C for 30 s, annealing temperature according to (Dequigiovanni *et al.*, 2014; Dequigiovanni *et al.*, in press, Chapter 2) for 45 s, and 72 °C for 45 s. Then, a second step with 8 cycles at 94 °C for 30 s, 53 °C for 45 s and 72 °C for 45 s, with a final extension at 72 °C for 10 min.

The resulting PCR products were evaluated in 1.5% agarose gels stained with GelRed (Biotium). Fragment separation and detection were performed on an ABI Prism 3130xl capillary sequencer using GeneScan 500 Rox-labelled size standard (Applied Biosystems). DNA fragment sizes were determined using GeneMapper software (Applied Biosystems).

5.2.3. Genetic structure and genetic diversity analyses

Genetic structure was inferred using STRUCTURE software v. 2.3.4 (Pritchard *et al.*, 2000). The admixture ancestry model with correlated allele frequencies was implemented with ten independent runs for each *K* (from 1 to 10), using 500,000 MCMC repetitions and a 200,000 burn-in period. Evanno's ΔK ad hoc statistics (Evanno *et al.*, 2005) was used to determine the optimal *K* value. Permutations of the output of STRUCTURE analysis were calculated with CLUMPP (Cluster Matching and Permutation Program) v. 1.1.2b (Jakobsson and Rosenberg, 2007) using independent runs to obtain a consensus matrix.

In order to examine the relationships among individual samples and geographic groups, neighbor-joining (Saitou and Nei, 1987) dendrograms were built with *Phylip 3.5* (Felsenstein, 2005), based on Cavalli-Sforza and Edwards' (1967) Chord distance (D_{CE}) obtained with *MSA 4.05* (Dieringer and Schlötterer, 2003). Final trees were formatted using *MEGA* version 7 (Kumar *et al.*, 2016). The dispersion of samples in the multivariate genetic variation space was evaluated by principal coordinate analysis (PCoA) obtained with GenAlEx v. 6.5 (Peakall and Smouse, 2012). Dispersion of the geographic groups was also determined using PCoA. The genetic variation within and among geographic groups and

Structure groups were evaluated using "locus-by-locus" AMOVA with GenAlEx v. 6.5 (Peakall and Smouse, 2012). Wright's F_{ST} was used to estimate population differentiation and was also calculated using GenAlEx. In addition, a Mantel test was carried out with the *adegenet* (Jombart and Ahmed, 2011) for R (R Core Team, 2015) for the correlation between Cavalli-Sforza and Edwards' genetic distances and geographic distances (km). Significance was assessed by conducting 10,000 permutations. Germplasm bank accessions were not considered for the Mantel analysis.

Based on the geographic groups and the Structure groups we estimated the following parameters of genetic diversity: number of alleles per locus (A), allelic richness (A_R) (El Mousadik and Petit, 1996), observed (H_O) and expected heterozygosity (H_E), and Wright's (1965) inbreeding coefficient (F_{IS}) with significance based upon 1,000 bootstrap replicates. The apparent outcrossing rate ($\hat{t}_a$) was obtained by the F_{IS} for each population (Vencovski, 1994) so that $\hat{t}_a = (1-F_{IS})/(1+F_{IS})$. Measures of genetic diversity by loci (number of alleles per locus, expected and observed heterozygosity and inbreeding coefficient) were estimated with *adegenet* (Jombart and Ahmed, 2011) for R (R Core Team, 2015).

5.3. Results

The samples collected presented ample morphological variation. The fruits varied in color, shape and size (Figure 5.2). Fruits with cylindrical, oval, flattened, duck-bill, heart-shape and snake-heads shapes were observed. Information on the color of the fruit was obtained from plants bearing fruits at the time of the collection. About 35 % of the fruits observed presented different shades of red, 35 % shades of green, 22 % shades of yellow and 8 % shades of green at the base and red at the tips.

Figure 5.2. Annatto (*Bixa orellana*) variability in fruit color and shape observed in samples collected for this study in Brazilian Amazonia

5.3.1. Genetic diversity

The 16 SSR loci amplified clearly and polymorphic peaks were easily scored in all genotypes. The total number of alleles found was 192, with an average of 12 alleles per locus (*A*). The number of alleles per locus varied widely among loci, ranging from a minimum of 4 (BorH7) to a maximum of 20 alleles (BorF9; Table 5.1). The effective number of alleles (*Ne*) varied between 1.268 (BorH7) and 4.149 (BorG11), with an average of 2.2. As indicated by the difference between the average value of *A* (12) and *Ne* (2.2), most alleles had frequencies lower than 0.05. Moderate mean observed (H_O=0.126) and moderate to high expected heterozygosities (H_E = 0.441) per loci were found. H_E values were higher than H_O, resulting in strong and positive inbreeding coefficient (F_{IS}) values, indicating an excess of homozygosity across the 16 loci (Table 5.1). Significant inbreeding coefficients, indicates deviations from Hardy-Weinberg equilibrium due to low frequency of heterozygotes. We estimated the mean apparent outcrossing rate as $\hat{t}_a$ = 0.198 and thus, the mean apparent selfing rate (0.802), indicates a mixed mating system for annatto, with predominance of inbreeding.

Table 5.1 Genetic parameters for the microsatellite analysis of *Bixa orellana* accessions with 16 SSR loci: *A*, number of alleles; H_O, observed heterozigosity; H_E, expected heterozigosity; F_{IS}, inbreeding coefficient

SSR Loci	A	H_O	H_E	F_{IS}
BorA2	6	0.192	0.584	0.651
BorA3	17	0.181	0.544	0.585
BorA5	9	0.109	0.507	0.801
BorB12	13	0.081	0.508	0.854
BorF9	14	0.158	0.361	0.546
BorG11	18	0.086	0.527	0.827
BorB1	14	0.194	0.538	0.580
BorB4	11	0.164	0.527	0.682
BorB10	15	0.057	0.231	0.776
BorC12	7	0.015	0.293	0.960
BorD1	20	0.414	0.694	0.403
BorD2	10	0.105	0.414	0.737
BorG4	19	0.133	0.755	0.822
BorH3	5	0.027	0.171	0.629
BorH7	4	0.029	0.180	0.780
BorH10	10	0.067	0.223	0.619
Overall Mean	12	0.126	0.441	0.703

Genetic diversity at the intragroup level was moderate to low. The mean number of alleles per locus ($\bar{A}$) ranged from 2.2 (São Paulo State, SP) to 7.2 (eastern Pará State), and the average number of alleles overall groups was 5.3 (Table 5.2). After adjusting for differences in sample sizes via rarefaction, allelic richness values did not differ substantially among the groups. Allelic richness (A_R) presented a mean value of 1.4, ranging from 1.3 (São Paulo State and Germplasm Bank) to 1.5. The mean number of effective alleles per locus (*Ne*) for all groups was 2.3 and the group from São Paulo State presented the lowest value (1.7). The mean H_O value for all groups was low (0.141). Gene diversity, inferred from Nei's heterozygosity (H_E), was moderate to high in all groups except for São Paulo State, which showed low levels of diversity (Table 5.2). Total gene diversity within the full sample was H_E = 0.462. Significant inbreeding coefficients (F_{IS}) were detected in all groups, with a mean of 0.669 (Table 5.2). We detected private alleles (61) in all but one group sampled (Table 5.2).

Table 5.2. Genetic diversity parameters in eight geographic groupings and four Stucture groups of *Bixa orellana* samples from Brazilian Amazonia, São Paulo and the Agronomic Institute's Annatto germplasm bank. [sample size (N), mean number of alleles ($\bar{A}$), allelic richness (Ar), mean number of effective alleles per locus (N_E), number of private alleles (Ap), observed (H_O) and expected (H_E) heterozygosity, and local inbreeding coefficient ($F_{IS} = 1 - H_O / H_E$)

Group	***N***	$\bar{A}$	***Ar***	***Ne***	***Ap***	H_O	H_E	F_{IS}
Rondônia State (12)	126	6.4	1.5	2.0	12	0.123	0.408	0.640
Madeira River (8)	66	5.3	1.5	2.7	8	0.157	0.549	0.721
Branco River (10)	53	6.7	1.5	2.4	10	0.159	0.512	0.691
Solimões River (6)	39	5.6	1.5	2.6	6	0.117	0.524	0.767
Amazonas River (6)	40	4.6	1.4	2.2	6	0.104	0.439	0.818
Pará State (17)	136	7.2	1.5	2.6	17	0.152	0.535	0.718
Germplasm Bank (2)	42	4.0	1.3	2.2	2	0.146	0.447	0.621
São Paulo State (0)	10	2.2	1.3	1.7	0	0.176	0.282	0.379
Average	-	5.3	1.4	2.3	7.6	0.141	0.462	0.669

5.3.2. Genetic structure of Brazilian annatto

Analysis of the population structure of all annatto accessions using STRUCTURE with Evanno *et al.*'s (2005) ΔK identified $K = 4$ as the most likely number of groups in the data set (Figure 5.2). However, $K = 2$ and 3 also revealed significant peaks.

At $K = 2$, Pará State, Germplasm bank accessions and São Paulo State samples were clustered separately from the other samples (Figure 5.3). At K=3, Pará State was classified in a separate group, and Rondônia State and Madeira River were also classified apart from the other major basins in Amazonia. At K=4, groups one (*yellow*), two (*gray*), three (*blue*) and four (orange) consisted of 121 (24 %), 110 (21 %), 138 (27 %) and 84 (17 %) annatto samples, respectively, with more than 60 % of their inferred ancestry derived from one of the respective groups (Figure 5.3). The remaining 59 samples (11 %) presented mixed ancestry.

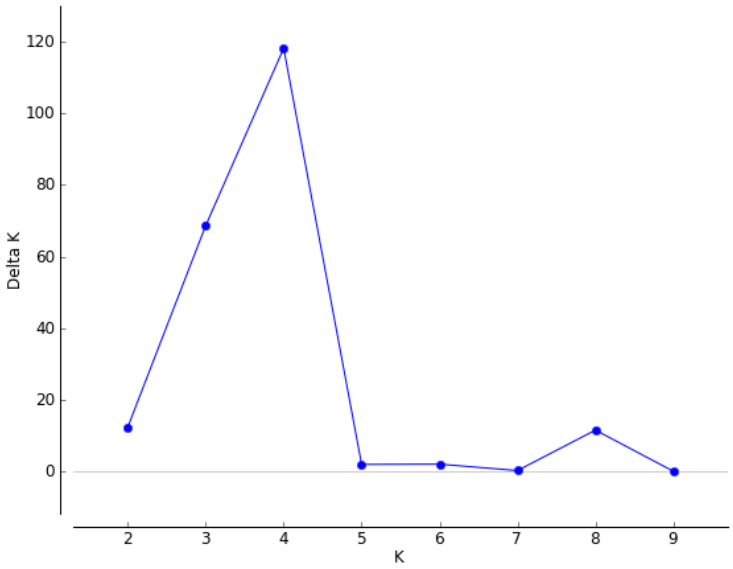

Figure 5.2 - Evanno *et al.* (2005) plot detecting the number of *K* groups that best fit the data for 16 SSR markers and 630 annatto accessions

Most of the samples from group 1 (yellow) originated in the state of Rondônia (70 accessions), and most of IAC germplasm bank accessions (34) are also in this group (Figure 5.3). The remaining samples are from the Solimões (5) and Madeira (5) rivers. Group 2 (gray) contained the Madeira River samples (53) and the remaining Rondônia samples (40). Also in this group are, in smaller amounts, samples from the Branco (9), Solimões (5) and Amazonas (2) rivers, and one from Pará. Group 3 (blue) contained most of the samples from the Amazonas (38), Branco (31) and Solimões (26) rivers. Ten samples from São Paulo State, 29 from Pará and three each from Rondônia and the Madeira River are also in this group. The geographical origins of all samples classified as group 4 (orange) were from eastern Pará (84).

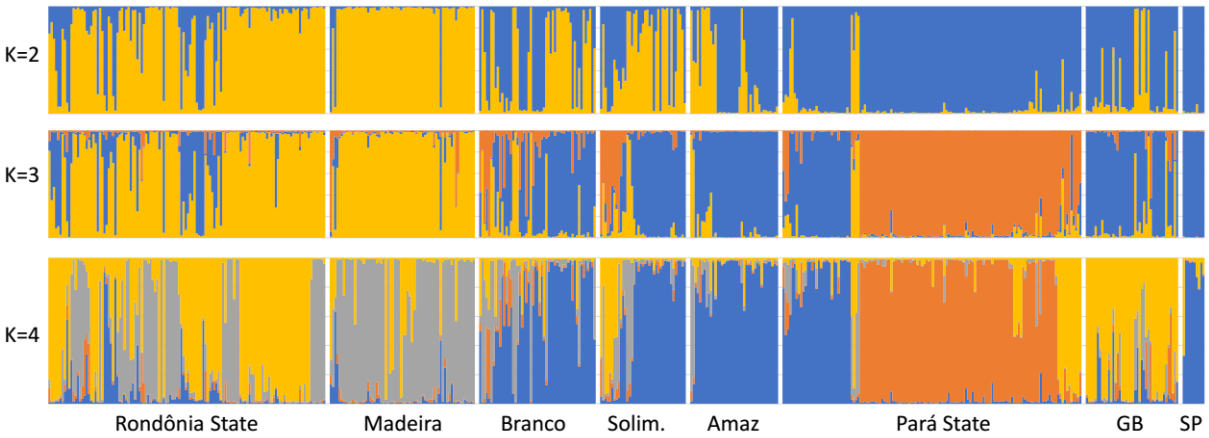

Figure 5.3 - Graphical display of sample assignment to the *K* = 2, 3 and 4 (most likely K) groups inferred with Structure based on 16 SSR loci and eight *Bixa orellana* geographic groupings, including the germplasm bank. Madeira = Madeira River, Branco = Branco and Negro Rivers; Solim. = Solimões River; Amaz. Amazonas River; GB = Germplasm bank; SP = São Paulo State

The first two principle coordinates accounted for 51.9 % of total genetic variability in the dataset (Figure 5.4A). There is a tendency for geographic structure across the PCoA space, but also admixture among groups. This is more evident when the geographic groups are planted (Figure 5.4B). In this analysis, the two principal coordinates accounted for 70.02 % of the variability. The São Paulo State and germplasm bank groups are most divergent. Corroborating the Bayesian analysis, the Rondônia and Madeira River groups are very close. Accessions from the Branco, Solimões and Amazon Rivers are also more related.

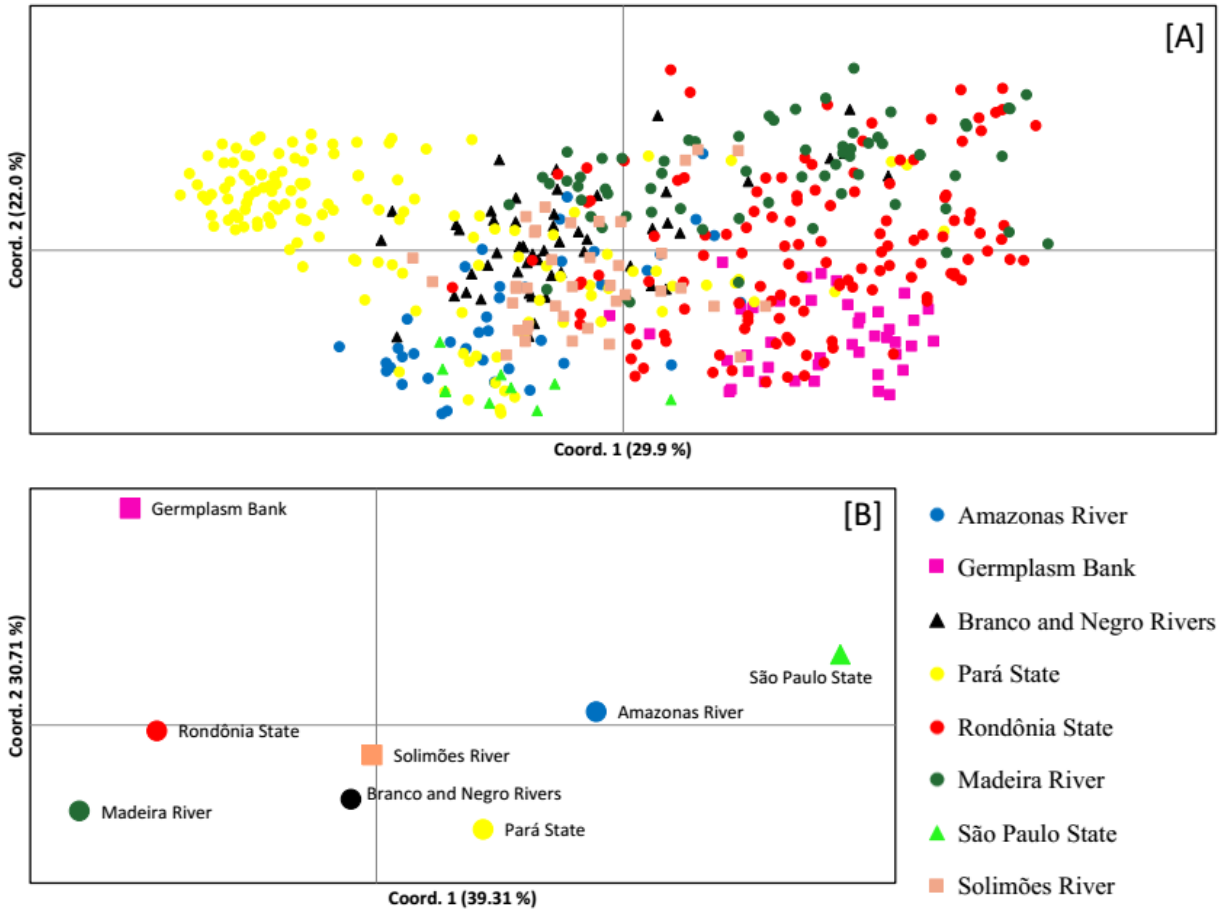

Figure 5.4 - Principal coordinate analysis (PCoA) based on Cavalli-Sforza and Edwards' (1967) chord distance [A] among the 512 accessions and [B] eight geographic groups of *Bixa orellana* and 16 SSR loci. Different symbols and colors represent the eight geographic groups

The neighbor-joining dendrograms (Figures 5.5 and 5.6) also showed a tendency to group annatto samples based on their geographic origin. In the NJ dendrogram of groups (Figure 5.6) the São Paulo and germplasm bank groups are more distant from the other groups. It also showed that Rondônia State samples are closer to the Madeira River samples. However, all our analyses demonstrated the high levels of admixture between groups.

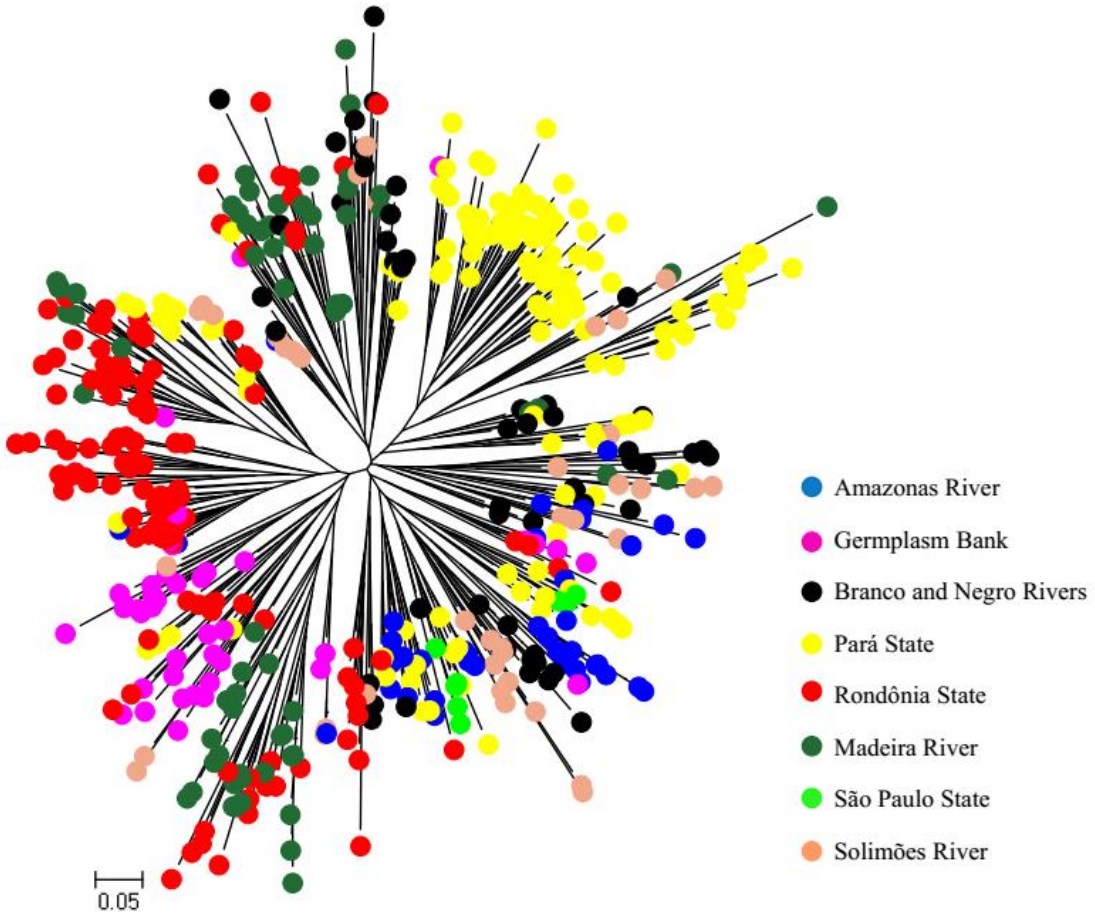

Figure 5.5 - Neighbor-joining dendrogram for individuals based on Cavalli-Sforza and Edwards' (1967) chord distance estimated from 16 nuclear SSR and 512 annatto (*Bixa orellana*) samples. Different colors represent the eight geographic groups

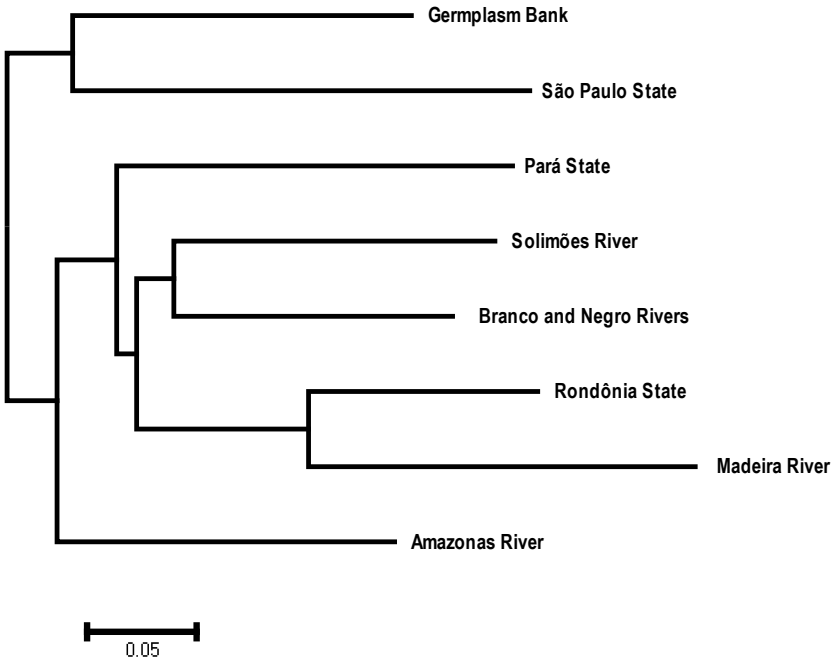

Figure 5.6 - Neighbor-joining dendrogram for geographic groups of annatto (*Bixa orellana*) based on Cavalli-Sforza and Edwards' (1967) chord distance estimated from 16 nuclear SSR

Genetic differentiation among most geographic groups was statistically significant ($p < 0.01$) (Table 5.3). F_{ST} values were moderate to low, ranging between 0.046 (Solimões – Amazonas) and 0.313 (São Paulo State – Madeira River). The largest F_{ST} values between groups were observed among São Paulo State and the other groups. The overall *F*-statistics estimates (F_{IS} = 0.743; F_{ST} = 0.188; F_{IT} = 0.787) confirmed high levels of genetic structure, showing that most of the diversity is concentrated within groups. Mantel's test showed a significant association (r = 0.525, p = 0.001) between pairwise genetic differentiation (*Fst*/(1 - *Fst*)) and geographic distance.

Table 5.3. Pairwise annatto (*Bixa orellana*) group estimates of F_{ST} (diagonal below) and probability based on 9,999 permutations (above diagonal)

Group	Rondônia	Madeira	Branco	Solimões	Amazonas	Pará	São Paulo
Rondônia	-	0.000	0.118	0.000	0.173	0.178	0.000
Madeira	0.081	-	0.103	0.000	0.199	0.185	0.000
Branco	0.000	0.000	-	0.000	0.092	0.000	0.000
Solimões	0.122	0.133	0.046	-	0.092	0.113	0.211
Amazonas	0.000	0.000	0.000	0.000	-	0.000	0.000
Pará	0.000	0.000	0.092	0.000	0.137	-	0.000
São Paulo	0.253	0.313	0.224	0.000	0.124	0.222	-

According to AMOVA, the within-group element explained most (86 %) of the total amount of genetic variation detected by our analyses. The remaining 14 % was due to differences between groups ($p < 0.001$) (Table 5.4).

Table 5.4. Analysis of Molecular Variance (AMOVA) performed for 16 SSR and 512 samples of cultivated annatto

Source	Degrees of freedom	Sum of Squares	Mean Square	Variance	%
Among Groups	6	594.395	99.066	0.757	14%
Within Groups	933	4205.950	4.508	4.508	86%
Total	939	4800.345	-	5.265	100%

5.4. Discussion

We observed moderate to high levels of admixture between geographic groups of cultivated annatto. The propagation of annatto is carried out, predominantly, by seed. It can also be performed by vegetative propagation (Pereira, 1995). However, the farmers visited do not use other forms of planting other than seeds. The plants obtained through sexual propagation present variability as a consequence of the genetic constitution, segregation and recombination between the genes, thus resulting in heterogeneous progenies (Lima, 1992). In

addition, farmers usually grow plants resulting from the mixing of native seeds and seeds coming from different regions of the country. In general, the regions of Rondônia and Pará are inhabited by migrants from several other states of the country. When farmers move from their hometown to another region, they usually carry samples of the plants they cultivated in their homegardens to plant in their new home. This practice results in the mixing of genotypes from different regions of the country (Mazzani *et al.*, 2000), as observed in the present study. In addition, a common practice observed is the exchange of seeds among farmers, resulting in more admixture. Therefore, our results highlighted a relevant importance of farmers' management practices able to influence the genetic structure of annatto, in particular the seed exchanges and selection.

Moderate levels of genetic diversity were observed among the nine annatto groups, highlighted by the mean number of alleles per locus ($\bar{A}$ = 5.3) and the mean observed (H_O = 0.141) and expected heterozygosity (H_E = 0.462). Similar results were obtained by Dequigiovanni *et al.* (submitted; Chapter 4), examining 63 accessions from the IAC germplasm bank of cultivated annatto. When comparing to a study conducted with 170 samples of wild annatto in Brazilian Amazonia ($\bar{A}$ = 3.856, H_O = 0.385, H_E = 0.519), higher mean number of alleles per locus was found. The process of domestication has significant consequences on crops, where domesticated populations show reduced genetic diversity in relation to the wild ancestor across the genome (Flint-Garcia, 2013). In comparison to the wild populations, domesticated populations have experienced a domestication bottleneck and generally successive bottlenecks due to distribution and diversification (Meyer and Purugganan, 2013). Crop populations also experience different selection pressures for greater yield, adaptation and quality (Gepts, 2004; Meyer and Purugganan, 2013).

High levels of inbreeding coefficient were observed suggesting deviations of HW equilibrium in all groups sampled. In other studies on annatto, the deviations from HW equilibrium using microsatellite markers were also reported (Dequigiovanni *et al.*, 2014, Chapters 2 and 3). When plants from different populations are grouped and analyzed in a single group, the expected heterozygosity is usually greater than the observed heterozygosity, which is known as the Wahlund effect (Hartl and Clark, 1998). Besides, experimental error, population sub-structuring, non-random crosses, selection and genetic drift were often invoked to explain the observed homozygote excess or deficit in allogamous species (Pressoir and Berthaud, 2004). Even though experimental error could not be excluded, Wahlund effect, inbreeding, genetic drift and the selection practiced by farmers are factors that might explain the excess of homozygotes observed in this study. The general excess of homozygotes

observed is also consistent with the results of previous studies, where different molecular and biochemical markers were used in the characterization of *B. orellana* (Carvalho *et al.*, 2005; Valdez-Ojeda *et al.*, 2010; Dequigiovanni *et al.*, 2014; Dequigiovanni *et al.*, in press, Chapters 2 and 4).

Bixa orellana is considered to be a cross-pollinated crop, due to its variability in morphological traits (Carvalho *et al.*, 2005) and studies with molecular markers, also reproducing through self-fertilization (Rivera-Madrid *et al.*, 2006; Joseph *et al.*, 2012). Dequigiovanni *et al.* (Chapter 6) estimated that the outcrossing rate in a cultivated annatto population measured with progeny tests and SSR markers (*Tm* = 0.532) is consistent with a mixed mating system. Thus, we can infer that the low $\hat{t}_a$ identified among the populations studied here is mainly due to the low occurrence of heterozygotes in the populations and may not necessarily be related to self-fertilization. Because cultivated fields are managed by the farmers, in some cases plants from different origins are grouped in the same plantation and analyzed in a single group. In these cases, the Wahlund effect is also observed (Hartl and Clark, 1998).

Levels of genetic diversity detectable in crops with molecular markers, such as microsatellites, are largely dependent on the mating system, the domestication history, and the magnitude of the collection being analyzed (Falk and Holsinger, 1991). SSR markers are important tools to assess genetic diversity of plant species (Govindaraj *et al.*, 2015). Usually, a relatively low number of SSR loci is sufficient to accurately reflect genetic structure and diversity among a high number of samples, as in the present study. The sampling used here is representative of Brazilian Amazonia, which is expected to be the center of origin of the crop (Clement *et al.*, 2016), as well as including samples maintained by the IAC germplasm bank of annatto and from the city of Piracicaba, in the State of São Paulo. This group from São Paulo, in addition to the absence of private alleles, presented very low rates of diversity compared to other groups. One of the reasons may be the low sampling of this group, with only 10 individuals, and all of them collected in one municipality, differing from the other groups, which could indicate a founder effect. The samples were collected in urban areas where annatto is used as an ornamental plant in the city of Piracicaba, São Paulo State. This material probably comes from nurseries and appears to be highly related. Also, it showed higher similarity to the samples from the Amazon River basin, as shown in the Bayesian, cluster and PCoA analyses. Annatto was traditionally used by indigenous people living in the Amazon Basin. Early nomadic hunter-gatherer populations must have dispersed annatto seeds

within the Amazon region. Later, annatto seed was probably also traded among the indigenous tribes contributing to an even wider dispersion of the species (Carvalho *et al.*, 2005), which could explain the proximity with the accessions from São Paulo.

A significant positive correlation between genetic and geographic distances according to Mantel's test indicates that annatto plants may be experiencing isolation by distance. The Bayesian analysis showed a higher similarity between accessions from Rondônia and the Madeira river accessions, having in mind that most of the accessions from Madeira river are located in the State of Rondônia, although a few accessions are also located in Amazonas State (Figure 5.1). This genetic proximity, examined on Groups 1 and 2 (at $K = 4$) of Figure 5.3, are probably mostly explained by the geographical proximity, with a few exceptions. The next group in the Bayesian analysis at $K = 4$ (Group 3) contained accessions from the main rivers of the Amazon basin, such as Amazonas, Branco, Negro and Solimões, clustered together due to genetic similarities and most probably also due to geographic proximity or/and interconnection between rivers, considering that this may facilitate the exchange of annatto varieties among riverside community farmers. And finally, the geographically most distant accessions from farmer fields in East Pará were clustered in a different group (Group IV, at $K = 4$), again corroborating the positive and significant correlation between genetics and geographical distances. Similar results were found in other studies with microsatellite markers in wild annatto (Chapter 3) and in a germplasm bank of annatto (Chapter 4). Valdez-Ojeda *et al.* (2008) using SRAP markers also found that annatto accessions from Yucatan, Mexico are clustered by geographic region.

5.5. Concluding remarks

The conservation of crop genetic resources is recognized as a key issue for biodiversity and for the resilience of agricultural systems (Frison *et al.*, 2011). To address the instability resulting from global environmental changes, many studies have stressed the importance of maintaining a wider genetic basis for cultivated plants (Hajjar *et al.*, 2008). We observed in this study moderate to high levels of admixture in cultivated annatto, suggesting gene flow between geographic groups, mainly by exchange of seeds among farmers. We also observed a decrease of observed heterozygosity in cultivated annatto, suggesting excess of homozygosity due to inbreeding and selection by farmers. Conservation strategies should be used to conserve this diversity. *In situ* conservation allows for a continued adaptation of plants

to the environment in which they occur (Altieri, 1999). In crop plants, *in situ* conservation is particularly important in areas under traditional agriculture, where crops are often enriched by gene exchange with wild relatives contributing to the long-term stability of agroecosystems (Altieri and Merrick, 1987). Traditional agroecosystems represent centuries of accumulated experience in interaction of farmers with the environment, using locally available resources, and often translated into systems with sustained yields. Many traditional agroecosystems are located in centers of diversity of crops, thus containing populations of variable and adapted landraces as well as wild relatives of crops (Harlan, 1992). The maintenance of traditional agroecosystems is the most indicated strategy to preserve *in situ* repositories of crop germplasm, including our study species, *B. orellana*. Therefore, we suggest that efforts be made to preserve the existing variability of annatto in the riverside communities along the main rivers of the Brazilian Amazonia, as well as in the State of Pará, aiming to maintain this diversity using strategies of *in situ* conservation, since these areas present more diversity besides presenting a high number of private alleles.

REFERENCES

Akshatha, V., Giridhar, P. and Ravishankar, G.A. (2011) Food , Ethanobotanical and Diversified Applications of *Bixa Orellana* L .: a Scope for Its Improvement Through Biotechnological Mediation. *Indian J. Fundam. Appl. Life Sci.* **1**, 9–31.

Altieri, M.A. (1999) The ecological role of biodiversity in agroecosystems. *Agric. Ecosyst. Environ.* **74**, 19–31.

Altieri, M.A. and Merrick, L.C. (1987) In situ conservation of Crop Genetic Resources through Maintentenance of Traditional Farming Systems. *Source Econ. Bot.* **41**, 86–96.

Andres, C. (1980) What are alternatives to use of FD & C yellow 5 in food? *Food Process.* **41**, 48–49.

Arce, J. (1999) El achiote *Bixa orellana* L. cultivo promisorio para el trópico. In Earth., p. 149.

Baleé, W. (2000) Antiquity of Traditional Ethnobiological Knowledge in Amazonia: The Tupi-Guarani Family and Time. *Ethnohistory* **47**, 399–422.

Cáceres, A., Menéndez, H., Méndez, E., Cohobón, E., Samayoa, B.E., Jauregui, E., Peralta, E. and Carrillo, G. (1995) Antigonorrhoeal activity of plants used in Guatemala for the treatment of sexually transmitted diseases. *J. Ethnopharmacol.* **48**, 85–88.

Carvalho, J.F.R.P. de, Robinson, I.P. and Alfenas, A.C. (2005) Isozymic variability in a Brazilian collection of annatto (Bixa orellana L.). *Pesqui. Agropecuária Bras.* **40**, 653–660.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967) Phylogenetic Analysis Models and Estimation Procedures. *Am. J. Hum. Genet.* **19**, 233–57.

Clement, C.R., Cristo-Araújo, M. de, D'Eeckenbrugge, G.C., Alves Pereira, A. and Picanço-Rodrigues, D. (2010) Origin and Domestication of Native Amazonian Crops. *Diversity* **2**, 72–106.

Clement, C.R., Rodrigues, D.P., Alves-Pereira, A., Mühlen, G.S., Cristo-Araújo, M. De, Moreira, P.A., Lins, J. and Reis, V.M. (2016) Crop domestication in the upper Madeira River basin. *Bol. do Mus. Para. Emilio GoeldiCiencias Humanas* **11**, 193–205.

Coelho, A.M.S.P., Silva, G.A., Vieira, O.M.C. and Chavasco, J.. (2003) Atividade antimicrobiana de *Bixa orellana* L. (Urucum). *Rev. Lecta* **21**, 47–54.

Dequigiovanni, G., Ramos, S.L.F., Zucchi, M.I., Bajay, M.M., Pinheiro, J.B., Fabri, E.G., Bressan, E.A. and Veasey, E.A. (2014) Isolation and characterization of microsatellite loci for *Bixa orellana*, an important source of natural dyes. *Genet. Mol. Res.* **13**.

Dieringer, D. and Schlötterer, C. (2003) microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* **3**, 167–169.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of Plant DNA from fresh tissue. *Focus (Madison).* **12**, 13–15.

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**, 2611–20.

Falk, D.A. and Holsinger, K.E. (1991) Genetics and conservation of rare plants, Oxford University Press.

Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6.

Flint-Garcia, S.A. (2013) Genetics and Consequences of Crop Domestication. *J. Agric. Food Chem.* **61**, 8267–8276.

Franco, C.F. de O., Fabri, E.G., Barreiro Neto, M., Manfiolli, M.H., Harder, M.N.C. and Rucker, N.C. de A. (2008) Urucum: sistemas de produção para o Brasil, João Pessoa : EMEPA-PB.

Frison, E.A., Cherfas, J. and Hodgkin, T. (2011) Agricultural Biodiversity Is Essential for a Sustainable Improvement in Food and Nutrition Security. *Sustainability* **3**, 238–253.

Gepts, P. (2004) Crop Domestication as a Long Term Selection Experiment, Oxford, UK: John Wiley & Sons, Inc.

Giuliano, G., Al-Babili, S. and Lintig, J. von (2003) Carotenoid oxygenases: cleave it or leave it. *Trends Plant Sci.* **8**, 145–9.

Giuliano, G., Rosati, C. and Bramley, P.M. (2003) To dye or not to dye: biochemistry of annatto unveiled. *Trends Biotechnol.* **21**, 513–6.

Govindaraj, M., Vetriventhan, M. and Srinivasan, M. (2015) Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives. *Genet. Res. Int.* **2015**, 431487.

Hajjar, R., Jarvis, D.I. and Gemmill-Herren, B. (2008) The utility of crop genetic diversity in maintaining ecosystem services. *Agric. Ecosyst. Environ.* **123**, 261–270.

Harlan, J. (1992) Crops and man, Madison: American Society of Agronomy.

Hartl, D. and Clark, A. (1998) Principles of Population Genetics, {Sinauer Associates}.

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. *Bioinformatics* **23**, 1801–1806.

Jombart, T. and Ahmed, I. (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071.

Joseph, N., Siril, E.A. and Nair, G.M. (2012) Reproductive characterization and preliminary studies on controlled breeding of Annatto (*Bixa orellana* L.). *Plant Syst. Evol.* **298**, 239–250.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R. and Dhawan, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177**, 309–334.

Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874.

Leal, F. and Clavijo, C.M. (2010) Acerca de la história, taxonomia, botánica y usos de *Bixa orellana* L. *Rev. Unell. Cienc. Tec* **1**, 78–86.

Lima, L. (1992) Opções incrementais sistematizadas na propagação da cultura do urucu (*Bixa orellana* L.). *Rev. Bras. Corantes Nat.* **1**, 223–226.

Lombello, R.A. and Pinto-Maglio, C.A.F. (2014) Cytogenetics and Reproductive Biology of *Bixa orellana* L. (Bixaceae). *Cytologia (Tokyo).* **79**, 379–386.

Mazzani, E., Marin, C.. and Segovia, V. (2000) Estudio de la variabilidad existente en la colección de onoto (Bixa orellana L.) del CENIAP; FONAIAP; Venezuela. *Rev. Fac. Agron.* **17**, 492–504.

Meyer, R.S. and Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852.

Morton, J.F. (1960) Can Annato (*Bixa orellana* L.) an old source of food color, meet new needs for safe dye. *Florida State Hortic. Soc.*, 301–309.

Mousadik, A. El and Petit, R.J. (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theor. Appl. Genet.* **92**, 832–839.

Nisar, N., Li, L., Lu, S., Khin, N.C. and Pogson, B.J. (2015) Carotenoid metabolism in plants. *Mol. Plant* **8**, 68–82.

Ohri, D., Bhargava, A. and Chatterjee, A. (2004) Nuclear DNA Amounts in 112 Species of Tropical Hardwoods - New Estimates. *Plant Biol.* **6**, 555–561.

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.

Pereira, T. (1995) Caracterização de plântulas de *Bixa orellana* L. - urucu (Bixaceae). *Rev. Bras. Sementes* **17**, 243–248.

Plotkin, M. (1993) Tales of a Shaman's apprentice, New York: Penguin Books.

Pressoir, G. and Berthaud, J. (2004) Patterns of population structure in maize landraces from the Central Valleys of Oaxaca in Mexico. *Heredity (Edinb).* **92**, 88–94.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

R Core Team (2015) R: A language and environment for statistical computing.

Rivera-Madrid, R., Escobedo-GM, R.M.M., Balam-Galera, E., Vera-Ku, M. and Harries, H. (2006) Preliminary studies toward genetic improvement of annatto (*Bixa orellana* L.). *Sci. Hortic. (Amsterdam).* **109**, 165–172.

Russell, K.R.M., Morrison, E.Y.S.A. and Ragoobirsingh, D. (2005) The effect of annatto on insulin binding properties in the dog. *Phyther. Res.* **19**, 433–436.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25.

Sandy-Cuen, P.M. and Becerra, R. (2003) Manejo campesino de recursos naturales. El achiote. *BioDiversitas* **7**, 7–11.

Schuelke, M. (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18**, 233–234.

Valdez-Ojeda, R., Hernandez-Stefanoni, J.L., Aguilar-Espinosa, M., Rivera-Madrid, R., Ortiz, R. and Quiros, C.F. (2008) Assessing morphological and genetic variation in Annatto (*Bixa orellana* L.) by sequence-related amplified polymorphism and cluster analysis. *HortScience* **43**, 2013–2017.

Valdez-Ojeda, R., Quiros, C.F., Lourdes Aguilar-Espinosa, M. de, Rivera-Madrid, R., Aguilar-Espinosa, M. de L. and Rivera-Madrid, R. (2010) Outcrossing rates in annatto determined by sequence-related amplified polymorphism. *Agron. J.* **102**, 1340–1345.

Vencovski, R. (1994) Variance of an estmatve of the outcrossing rate. *Rev. Bras. Genética* **17**, 349–351.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. (2016) Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **39**, 312–328.

Wright, S. (1965) The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution (N. Y).* **19**, 395.

Yolmeh, M., Habibi Najafi, M.B. and Farhoosh, R. (2014) Optimisation of ultrasound-assisted extraction of natural pigment from annatto seeds by response surface methodology (RSM). *Food Chem.* **155**, 319–324.

6. MIXED MATING SYSTEM OF ANNATO (*BIXA ORELLANA*) DETERMINED WITH MICROSATELLITE MARKERS

Abstract

Knowledge of the breeding systems of crop plants is important for plant breeding and conservation. Using 12 microsatellite loci, a *Bixa orellana* population from Rondon do Pará, in the State of Pará, Brazil, was evaluated to determine its breeding system. This population, collected in a field of a local variety of annatto, was represented by 20 families with 15 individuals each. Microsatellite fingerprints were obtained with an ABI Prism 3130xl capillary sequencer using GeneScan 500 Rox-labelled size standard. Mating system analysis was based on the mixed mating and correlated mating models using the multilocus mating system program MLTR. Multilocus outcrossing rate was 0.532, indicating a mixed mating system for this *B. orellana* population. Biparental inbreeding (0.081) contributed to the selfing rate in this population. The paternity correlation was high (0.260), indicating that more than 25 % of the outcrossing occurred between related individuals. The parental fixation index (*F*) was estimated at 0.033, indicating the presence of inbreeding. As a result of selfing and high paternity correlation, the coancestry coefficient within families (0.313) was higher than expected in half sib progenies (0.125). The estimated effective number of pollen donors in this population was 3.8, meaning that approximately 4 males crossed to each of the 20 females. In conclusion, due to this mixed breeding system, the collection of open-pollinated seeds for plant breeding or conservation purposes should include at least 100 trees in order to ensure a representative sample.

Keywords: Outcrossing; Inbreeding; SSR; Amazonia; Breeding System

6.1. Introduction

Mating systems and reproductive biology plays a key role in the demographic process, determining population genetic structure and diversity, influencing in the effective population sizes and population subdivisions (Ward *et al.*, 2005). The effect of the mating system on genetic diversity is fundamental in plant evolutionary genetics, as gene flow is primordial in structuring the genetic variability within and among populations (Barnaud *et al.*, 2008). Domesticated plants exhibit several types of mating systems, ranging from obligatory outcrossing to self- fertilization and, in some cases, apomixis (Ross-Ibarra *et al.*, 2007). Selfing promotes population differentiation, local adaptation, and in some cases it can result in the expression of recessive and deleterious alleles. On the other hand, random outcrossing promotes gene flow, increases heterozigosity, and in association with other evolutionary factors homogenizes populations (del Castillo and Trujillo, 2007). Gene flow requires not only outcrossing but also the incorporation of new genetic combinations in the population

(Slatkin, 1987). In domesticated plants, incorporation of new genetic combinations may involve natural selection and human selection exerted by farmers.

In traditional farming systems, biological and human factors interact to shape evolutionary forces (Alvarez *et al.*, 2005). Biological factors comprise both environmental pressures and biological traits of plants, including their mating system. Human factors affect the dynamics of diversity in many ways, acting on gene flow, drift and selection (Altieri, 1999). Mating system is one of the characteristics of domestic species that can be modified by human selection (Meyer *et al.*, 2012). Most trees, including small trees or shrubs, have allogamy as their main strategy (Ward *et al.*, 2005).

Estimates of outcrossing rates and the number of pollen donors within and among progenies of populations can help to elucidate the genetic structure of these populations and contribute in studies of the inheritance of quantitative traits, selection of superior genotypes, and determination of sample sizes for germplasm collections and conservation strategies (Cuthbert and McVetty, 2001; Karasawa *et al.*, 2007). Strategies for conservation are becoming more complex and difficult to manage as climate change accelerates during this century, and information on the mating system is essential to contribute to this conservation effort (Jarvis *et al.*, 2008).

Bixa orellana L. (known as ‘annatto’, ‘achiote’ or ‘urucum’) is a perennial crop native to the Neotropics (Arce, 1999; Rivera-Madrid *et al.*, 2006), more specifically from the Amazon region (Sandy-Cuen and Becerra, 2003; Clement *et al.*, 2010; Moreira *et al.*, 2015; Clement *et al.*, 2016;). Annatto is commercially valuable due to its seeds, which contain large amounts of a specific reddish-orange pigment, bixin, an apocarotenoid widely used as a colorant in food, and cosmetics industries (Nisar *et al.*, 2015). Besides the food-industrial use, studies have pointed out some medicinal properties in addition to dye, such as geranylgeraniol, tocotrienols and other carotenoids presenting antimicrobial, antioxidant and antiviral properties (Cáceres *et al.*, 1995; Coelho *et al.*, 2003; Yolmeh *et al.*, 2014).

Bixa orellana is distributed over a wide geographic scale and adapted to a broad range of edaphic and climatic conditions (Joseph and Siril, 2014). Mainly propagated by seeds (Rivera-Madrid *et al.*, 2006), this crop is diploid with 2n = 14 chromosomes (Lombello and Pinto-maglio, 2014). Although outcrossing predominates in the species, self-pollination may occur. Almeida & Pinheiro (1992) described a series of characteristics of the floral biology in the “verde piloso” variety of annatto with controlled pollinations and concluded that cross-fertilization is the method for reproduction and natural autogamy doesn’t occur in this annatto type. Rivera-Madrid *et al.* (2006), also carrying out controlled pollination and floral biology

studies in accessions from Mexico, obtained intermediate levels of cross-pollination. Valdez-Ojeda *et al.* (2010), analyzing eight progenies of self-pollination with 10 individuals with SRAP (Sequence-Related Amplified Polymorphism) markers, also in Mexico, concluded that the species reproduces by cross-pollination. Joseph *et al.* (2012) carried out very detailed studies on the reproductive biology of annatto in India, also concluding that the species presents cross-pollination and may also reproduce by self-fertilization.

However, none of these studies evaluated progenies of open fertilization, and neither from the area where the crop originated, in Amazonia. The aim of this study was to evaluate the annatto mating system in an open-pollinated population from Rondon do Pará, PA, in the Brazilian Amazonia, through progeny tests. We used microsatellite markers or simple sequence repeats (SSR) that are considered ideal for studying mating systems because they feature hypervariability and co-dominance, and besides presenting highly reproducible results, exhibiting simple Mendelian inheritance and are widely distributed in the nuclear genome (Kalia *et al.*, 2011; Vieira *et al.*, 2016). Studies on mating systems using microsatellites have been conducted for several Neotropical species (Ward *et al.*, 2005; Ramos *et al.*, 2011; Abreu *et al.*, 2012; Nazareno and Reis, 2012; Picanço-Rodrigues *et al.*, 2015)

Our main objective was to investigate the mating system of *Bixa orellana* in order to identify the levels of cross-fertilization and/or self-fertilization, and to better understand the genetic structure of progenies in a population of cultivated annatto from Brazilian Amazonia, providing information for conducting and establishing conservation and breeding programs. Our study addressed the following specific questions: (1) What is the outcrossing and paternity correlation rate in this population? (2) Is there any mating among relatives in this population? And (3) what are the effective number of pollen donors, the average co-ancestry coefficient and the variance effective size within progenies?

6.2. Material and methods

6.2.1. Plant material

The mating system study of annatto was conducted based in a population located in Rondon do Pará – PA, Brazil (4°44'43.5"S; 47°55'59.4"W). This population was maintained by a traditional farmer. In this property, seeds from 20 mother plants were collected. From each plant, around six fruits were randomly collected and from each fruit 20 seeds were collected. The seeds were submitted to a dormancy break treatment by submersion in water

for 24 hours. The seeds were grown under greenhouse conditions with soil and Agrolite (2:1). From each progeny, 15 seedlings were selected for DNA extraction and analysis of the reproductive system.

From each seedling, two to four leaves were ground in liquid N_2 and total genomic DNA was isolated following Doyle and Doyle (1990) with minor modifications. DNA concentration was estimated by comparison with phage λ (lambda DNA—Invitrogen) in agarose gels (1%) electrophoresis stained with GelRed (Biotium). Microsatellite amplifications were performed using 12 microsatellite markers (BorA2, BorA3_2013, BorA5_2013, BorB1_2013, BorB4, BorB5_2013, BorB12_2013, BorC5_2013, BorD1_2013, BorF9_2013, BorG11_2013 and BorH10_2013) developed for *Bixa orellana* (Dequigiovanni *et al.* 2014; Dequigiovanni *et al.* in press) selected based on polymorphism and multiplexing capability. The amplifications were performed in an a MyCycler Thermal Cycler (Bio-Rad, Hercules, CA, USA) according to the conditions described in Dequigiovanni *et al.* (2014) and Dequigiovanni *et al.* (in press, Chapter 2). Microsatellite data was obtained automatically with an ABI Prism 3130xl capillary sequencer using GeneScan 500 Rox-labelled size standard.

6.2.2. Data analysis

This study was based on the mixed mating (Ritland and Jain, 1981) and correlated mating models (Ritland, 1989; Ritland, 2002) implemented in MLTR - multilocus mating system program (Ritland, 2004). Mating system parameters were estimated using the expectation maximization method, which is recommended for data sets with missing data, small number of progenies, and allowing the presence of undetected null alleles (Ritland, 2002). Therefore, the following parameters at the population level were estimated: Multilocus outcrossing rate (t_m); Single locus outcrossing rate (t_s); Fixation index of maternal parents (F); Correlation of paternity (r_p); Correlation of selfing (r_s); and the correlation of t among loci ($r_t(l)$). Multilocus outcrossing rate (t_m) was also estimated at progeny level. The difference $t_m - t_s$ was used to estimate biparental inbreeding, the increase in homozygosity resulting from mating between relatives. Under biparental inbreeding, the magnitude of this difference should be positive, as single locus estimates of outcrossing rates will include apparent selfing due to mating between relatives (Ritland, 2002). Standard errors of the estimates were approximated as the standard deviation of 10,000 bootstraps replicates, resampling maternal families and individuals.

The co-ancestry coefficient among plants within progenies (θ_{xy}) was estimated from the correlation coefficient of relatedness among plants within progenies (r_{xy}), as proposed by Ritland (1989). In non-inbred diploids, the co-ancestry coefficient is half of r_{xy}, so that $\theta_{xy}=r_{xy}/2$. The co-ancestry within families was defined as:

$$\theta = 0.125(1 + \hat{F}_a)\left[4\hat{s} + (\hat{t}_m^2 + \hat{t}_m\hat{s}\hat{r}_s)(1 + \hat{r}_{p(m)})\right] \quad .$$

where, $\hat{s}$ is the self-fertilization rate obtained from $\hat{s} = 1 - \hat{t}_m$. $\hat{F}_a$ is the inbreeding coefficient in the parental population (Ritland, 1989). The variance effective population size was estimated following the expression: $N_{e(v)}= 0.5/\theta$ (Cockerham, 1969). The average effective number of pollen donors per seed tree, was estimated by: $N_{e(p)} = \left[1/\hat{r}_{p(m)}\right]$ (Ritland, 1989).

6.3. Results and discussion

6.3.1. Outcrossing rates

All 12 microsatellite loci used to estimate the parameters of the mating system of annatto were polymorphic. The multilocus outcrossing rate (t_m) based on the 12 loci was 0.532, suggesting that the species presents a mixed mating system with a slight predominance of cross-fertilization (Table 6.1).

Few studies have been reported so far on the mating system and outcrossing rates in *Bixa orellana*. The species is considered by its variability of contrasting morphological traits to be a cross-pollinated crop (Carvalho *et al.*, 2005), although previous studies have suggested that annatto can tolerate both pollination types (cross- and self-pollination). Rivera-Madrid *et al.* (2006), carrying out controlled pollinations in accessions from Mexico, observed 57% of cross-pollination and 31.4% of self-pollination, consistent with data obtained from the population of the State of Pará evaluated in this study (53% of cross-pollination). Valdez-Ojeda *et al.* (2010), analyzing eight progenies of self-pollination with 10 individuals each, using the dominant marker SRAP, also in Mexico, concluded that the species reproduces by cross-pollination and estimated a higher multilocus crossover rate of 0.748. Joseph *et al.* (2012), in a detailed study on the reproductive biology of annatto in India, also concluded that the species presents cross-pollination and may also reproduce by self-fertilization. Almeida

and Pinheiro (1992) described a series of characteristics of the floral biology of annatto, also performing controlled pollinations, suggesting that cross-fertilization is the natural method of reproduction in annatto. The data obtained in the present study also indicates the reproduction of annatto by cross-fertilization, but with a considerable level of self-fertilization.

Outcrossing can be promoted mainly because flower maturation of annatto has been shown to be asynchronous on the same panicle and between different panicles (Vallejo, 1991; Rivera-Madrid *et al.*, 2006; Akshatha *et al.*, 2011). Annatto has also been proved to be protandrous (Joseph *et al.*, 2012). Protandri is a prezygotic mechanism that favors outcrossing, reducing the frequency of selfing and biparental inbreeding (Charlesworth and Charlesworth, 1987) and occurs when the anthers shed their pollen before the stigma of the same flower is receptive (Rivera-Madrid *et al.*, 2006).

On the other hand, annatto presents hermaphrodite flowers (Morton, 1960) and the position of female and male sexual organs appears to promote selfing (Rivera and Flores, 1988). Lombello & Pinto-Maglio (2014) detected that pollen tubes reach the ovules in *in vitro* self-pollinations, showing that *B. orellana* is self-compatible. This fact also reinforces the existence of maturing synchrony between stigma and pollen proposed by Almeida & Pinheiro (1992). Another characteristic that favor selfing is that *B. orellana* is not totally dependent of pollinator agents, although the number of formed fruits and fertility rates significantly decrease with the absence of external pollination (Lombello and Pinto-maglio, 2014).

The individual multilocus outcrossing rates per progeny showed high variation among progenies, ranging from 0.143 to 1. This variation in outcrossing rates may be associated with the spatial isolation of some plants, since the sampled population is small and in many cases some sampled plants were at a higher distance from the other sampled mother plants. Spatial isolation may lead to reproductive isolation, causing pollinators to visit more flowers from the same plant, leading to self-fertilization. This patterns has also been observed in many mating system studies in tropical plants (Dick *et al.*, 2003; Ward *et al.*, 2005).

Alternatively, the variation in the rate of crossing between progenies may be associated with the occurrence of inbreeding depression. Some plants may contain more genetic load than others, so that self-fertilizations do not result in offspring due to embryo abortion. Floral morphology and self-compatibility system are also expected to influence the observed outcrossing rates in plant species (Barrett, 1998). Artificial pollination studies (Rivera-Madrid *et al.*, 2006) have indicated that crossing between plants with white flowers and plants with pink or red flowers resulted in a higher number of seeds per fruits than in crosses involving only white flowers. This suggests inbreeding depression with respect to

white flower variants because they are probably related by ascendance. Valdez-Ojeda *et al.* (2010), studing outcrossing rates in annatto determined by SRAP, observed high outcrossing rates for both types (white and pink flowers) but identified that white flowers exibited an excess of homozygotes, in contrast to pink flower variants that presented an excess of heterozygotes according to the Wright fixation index. Joseph & Siril (2013) also found that the distance between anther and stigma and the number of anthers per flower was significantly different in rose and purple flowers. Together with the position of stigma, these features can promote different levels of cross-pollination. They also concluded that floral color influences pollinator behavior in *B. orellana*. Pollinator behavior can result in changes in plant fertility, out-crossing rates, and pollen flow between plants, resulting in differential reproductive success (Devaux *et al.*, 2014). Petal color may give information to the pollinators regarding the quantity and quality of pollen associated with color (Joseph and Siril, 2013).

Table 6.1. Estimates of mating system parameters in progenies from an annatto (*Bixa orellana*) population cultivated in Rondon do Pará – PA, Brazil

Estimate	Outcrossing rate
	MME method
Progeny 1	0.484 (0.123) [15]
Progeny 2	0.530 (0.138) [15]
Progeny 3	0.421 (0.521) [15]
Progeny 4	0.622 (0.161) [15]
Progeny 5	0.885 (0.079) [15]
Progeny 6	0.306 (0.122) [15]
Progeny 7	0.518 (0.320) [15]
Progeny 8	0.143 (0.498) [15]
Progeny 9	0.614 (0.146) [15]
Progeny 10	0.461 (0.136) [15]
Progeny 11	0.798 (0.175) [15]
Progeny 12	0.647 (0.183) [15]
Progeny 13	1.000 (0.097) [15]
Progeny 14	0.537 (0.313) [15]
Progeny 15	0.897 (0.047) [15]
Progeny 16	0.545 (0.407) [15]
Progeny 17	0.418 (0.299) [15]
Progeny 18	0.192 (0.143) [15]
Progeny 19	1.000 (0.045) [15]
Progeny 20	1.000 (0.007) [15]
Multilocus outcrossing rate: t_m	0.532 (0.047)
Single locus outcrossing rate: t_s	0.451 (0.060)
Mating among relatives: $t_m - t_s$	0.081 (0.026)
Correlation of the estimate of t: r_t	0.231 (0.065)
Correlation of s among loci: r_s	0.860 (0.101)
Correlation of the estimate of p multilocus: r_p	0.260 (0.071)
Number of pollen donors: N_{ep}	3.846
Selfing rate: - ŝ= 1 - $\dot{t}_m$	0.468
Parental fixation index: F	0.033
Coancestry within progenies: Θ_{xy}	0.313
Variance effective size: $N_{e(v)}$	1.596

Caro *et al.* (2016) highlighted that annatto can produce a limited amount of fruit by self-pollination and cross-polination can substantially increase benefits in quantity and quality of fruits. Lombello & Pinto-maglio (2014) also observed that fruits formed exclusively by self-pollination were smaller and had fewer seeds than those obtained by open pollination, highlighting the occurence of inbreeding depression and the importance of pollination for annatto seed production. Hufford & Hamrick (2003) studying *Platipodiun elegans* also detected that depression due to inbreeding altered the crossing rate between fertilization and the seedling stage, when the genes of the seedlings were accessed by molecular analysis.

6.3.2. Mating among relatives

The average single-locus outcrossing rate (t_s) was 0.451 (0.060). Values of t_m and t_s differed significantly from each other (P<0.01). The difference between the multilocus outcrossing rate and single-locus outcroossing rate ($t_m - t_s$) has been used to quantify the occurrence of crossings among related individuals. In this study, this difference was 0.081 (0.026) suggesting that mating between closely related individuals is occurring in this population. For cultivated plants, there is a higher expectation of mating among relatives, as it is common to have only a few seed sources per cultivated area due to farmers' selection practices (Louette, 2000). Biparental inbreeding, or mating between relatives, apparently causes selfing or increased homozygosity (Ritland, 2002). Crossing between relatives occurs due to the existence of spatial genetic structure, caused mainly by the dispersion of seeds close to the mother plants, so that related individuals grow in a group. Associated with this, the pollinators behavior may favor the occurrence of crosses between nearby plants, especially if the population density of reproductive individuals is low due to the small size of the populations or individual variations in the flowering. Crossing among relatives were also described in many other studies with tropical species (Ward *et al.*, 2005).

6.3.3. Correlated matings

Correlated crosses were measured by multilocus paternity correlation (r_p), which measures the likelihood of two random individuals having the same pollen donor. The multilocus paternity correlation within progeny arrays was very high and significantly different from zero (r_p = 0.260, SE = 0.071), indicating correlated matings and suggest that more than 25% of the open-pollinated pairs of offspring are full-sibs. Correlated crosses can be generated by the behavior of the pollinators, fecundating the same flowers with pollen from the same plant or due to the small size of the population. Correlated crosses, together with self-fertilization rate, show that the open-pollinated progenies of almost all parent plants are composed of different types of relatedness, and may contain half-siblings, full-siblings and self-fertilization siblings.

The parental fixation index (F) was estimated at 0.033, indicating the presence of inbreeding. Wright's fixation index F provides an estimate of what happened historically in the population. Values close to zero suggest a history of outcrossing population. On the other hand, the outcrossing rate t is an indirect estimate of what has happened during the most

recent generation and is less affected by factors such as selection and drift than *F* is. Therefore, *F* and *t* values may seem to suggest different levels of outcrossing in a population, but they may not actually be contradictory (Cook and Soltis, 2000). Wright's fixation value in the progeny screened with SSR markers in this study was lower than expected, based on the estimate of t_m. With a t_m value of 0.532, the expected fixation index was [$F = (1 - t)/(1 + t)$] = 0.305, while the estimated value of *F* was 0.033. A lower than expected *F* suggests an excess of heterozygotes and less inbreeding than expected in the progeny suggesting a history of outcrossing in this population.

As a result of selfing and high paternity correlation, the coancestry coefficient within families (0.313) was much higher than expected in halfsib progenies (0.125), showing that progeny arrays are on average more related than half-sibs. The coancestry coefficient plays an important role in breeding programs and conservation strategies of plant genetic resources due to additive genetic variation and heritability estimation (Sebbenn *et al.*, 2000). According to Ritland (1989), in half-sib families, the relatedness coefficient estimates 25% of the additive genetic variation. Similar coancestry coefficients were found in other Neotropical species such as *Calophyllum brasiliense* (Brotel *et al.*, 2006) with coancestry coefficient of 0.289 and *Ilex paraguariensis* (Gomes *et al.*, 2009), that presented a coancestry coefficient of 0.253.

The mean number of individuals that effectively pollinate maternal plants may be estimated using the correlation of paternity, that is, the mean number of probable individuals that contributed pollen to the reproductive event of a maternal tree. In this study, the estimated effective number of pollen donors in this population was 3.8, meaning that approximately 4 males crossed to each of the 20 females. This estimate is considered to be a very low number and can result in a reduced diversity in the population. A higher number of pollen donors can help to increase the genetic diversity and variance effective population size within progenies (Picanço-Rodrigues *et al.*, 2015).

6.3.4. Effective population size

The variance effective size in this annatto population was N_{ev} = 1.592, requiring a minimum number of 60 seed-trees to maintain an effective size of 100. However, it is important to emphasize that seed plants for seed collection may not be related to one another, otherwise their progeny will be relatives, which reduces the effective size. In order to avoid collecting seeds from neighboring parent trees, we should avoid collecting seeds from nearby

spatially adjacent plants. The ideal is to collect seeds from plants located at great distances. Studies have shown that plant species pollinated by bees can receive pollen from long distances (> 1000 m), as reported for by *Miracrodruon urundeuva* (Gaino *et al.*, 2010), *Tabebuia aure* (Braga and Collevatti, 2011), and *Copaiffera langsdorffii* (Manoel *et al.*, 2012). However, the strongest the isolation effect by distance, the smaller the frequency of crosses between them. Costa *et al.* (2008) and Almeida & Pinheiro (1992) identified that the effective pollinators of annatto flowers are large bees of the species *Xylocopa frontalis*, *Eulaema nigrita* and *E.cingulata*. Lombello & Pinto-maglio (2014) also observe bumble bees (Bombus sp.) and ants visiting *B. orellana* flowers. However, as there are no studies related to the distances that these bees manage to carry pollen of annatto, a strategy could be the collection of seeds of trees located at a distance of at least 1,000 m.

6.4. Concluding remarks

After studying the mating system of the *B. orellana*, we can confirm that it has a mixed mating system. The mating system plays a crucial role in the amplification and recombination of the genetic variability of populations. Consequently, the random mating deviations observed in *B. orellana* have important consequences for conservation and breeding. *Ex situ* conservation of populations with deviations from panmixia demands larger samples than those recommended for outcrossing populations because these deviations cause the sample effective size to decrease. Our estimates of the number of plants necessary for *ex situ* conservation are extremely high and similar to most outcrossing and mixed mating tropical plants, which means that *ex situ* conservation may not be the best way to conserve this diversity.

Our findings also strongly indicate a need for studies concerning the nature of pollination and its relation to the mating system, besides the contribution of pollinating agents and seed dispersal, since evidence suggests that there are several characteristics influencing cross-pollination and that pollinators may play an important role in these processes. Analyzing the mating system of different cultivated populations, with different degrees of cultivation, may also be important in future studies, considering that higher cultivation pressures should lead to higher selfing rates expectations. The comparison of these studies with wild annatto populations should also bring interesting results, and the hypothesis is for higher outcrossing rates in wild populations. Future studies in different environments should

also be evaluated since outcrossing can be highly influenced by climatic conditions. These results need to be taken into account for breeding and conservation programs of annatto.

REFERENCES

Abreu, A.G., Priolli, R.H.G., Azevedo-Filho, J.A., Nucci, S.M., Zucchi, M.I., Coelho, R.M. and Colombo, C.A. (2012) The genetic structure and mating system of Acrocomia aculeata (Arecaceae). *Genet. Mol. Biol.* **35**, 119–121.

Akshatha, V., Giridhar, P. and Ravishankar, G.A. (2011) Food, Ethanobotanical and Diversified Applications of *Bixa orellana* L .: a Scope for Its Improvement Through Biotechnological Mediation. *Indian J. Fundam. Appl. Life Sci.* **1**, 9–31.

Almeida, E. and Pinheiro, A. (1992) Biologia floral e mecanismo de reprodução em urucuzeiro (*Bixa orellana* L.) I. tipo "fruto verde piloso." In Reunião técnico científica sobre melhoramento genético do urucuzeiro., pp. 72–81. Belém: Embrapa - CEPATU.

Altieri, M.A. (1999) The ecological role of biodiversity in agroecosystems. *Agric. Ecosyst. Environ.* **74**, 19–31.

Alvarez, N., Garine, E., Khasah, C., Dounias, E., Hossaert-Mckey, M. and McKey, D. (2005) Farmers' practices, metapopulation dynamics, and conservation of agricultural biodiversity on-farm: A case study of sorghum among the Duupa in sub-sahelian Cameroon. *Biol. Conserv.* **121**, 533–543.

Arce, J. (1999) El achiote *Bixa orellana* L. cultivo promisorio para el trópico. In Earth., p. 149.

Barnaud, A., Trigueros, G., McKey, D. and Joly, H.I. (2008) High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? *Heredity (Edinb).* **101**, 445–452.

Barrett, S.C.H. (1998) The evolution of mating strategies in flowering plants. *Trends Plant Sci.* **3**, 335–341.

Braga, A.C. and Collevatti, R.G. (2011) Temporal variation in pollen dispersal and breeding structure in a bee-pollinated Neotropical tree. *Heredity (Edinb).* **106**, 911–919.

Brotel, G., Marcos de Souza, A., Carvalho, D., Isabel do Pinto, S., Oliveira, M. and Abjaud, R. (2006) Caracterização genética de *Calophyllum brasiliense* Camb. em duas populações de mata ciliar. *Soc. Investig. Florestais* **30**, 821–827.

Cáceres, A., Menéndez, H., Méndez, E., Cohobón, E., Samayoa, B.E., Jauregui, E., Peralta, E. and Carrillo, G. (1995) Antigonorrhoeal activity of plants used in Guatemala for the treatment of sexually transmitted diseases. *J. Ethnopharmacol.* **48**, 85–88.

Caro, A., Moo-Valle, H., Alfaro, R. and Quezada-Euán, J.J.G. (2016) Pollination services of Africanized honey bees and native *Melipona beecheii* to buzz-pollinated annatto (*Bixa orellana* L.) in the neotropics. *Agric. For. Entomol.*

Carvalho, J.F.R.P. de, Robinson, I.P. and Alfenas, A.C. (2005) Isozymic variability in a Brazilian collection of annatto (*Bixa orellana* L.). *Pesqui. Agropecuária Bras.* **40**, 653–660.

Castillo, R.F. del and Trujillo, S. (2007) Effect of inbreeding depression on outcrossing rates among populations of a tropical pine. *New Phytol.* **0**, 071107070910001

Charlesworth, D. and Charlesworth, B. (1987) Inbreeding Depression and its Evolutionary Consequences. *Annu. Rev. Ecol. Syst.* **18**, 237–268.

Clement, C.R., Cristo-Araújo, M. de, D'Eeckenbrugge, G.C., Pereira, A.A. and Picanço-Rodrigues, D. (2010) Origin and domestication of native Amazonian crops. *Diversity* **2**, 72–106.

Clement, C.R., Rodrigues, D.P., Alves-Pereira, A., Mühlen, G.S., Cristo-Araújo, M. De, Moreira, P.A., Lins, J. and Reis, V.M. (2016) Crop domestication in the upper Madeira River basin. *Bol. do Mus. Para. Emilio Goeldi Ciencias Humanas* **11**, 193–205.

Cockerham, C.C. (1969) Variance of Gene Frequencies. *Evolution (N. Y).* **23**, 72.

Coelho, A.M.S.P., Silva, G.A., Vieira, O.M.C. and Chavasco, J.. (2003) Atividade antimicrobiana de *Bixa orellana* L. (Urucum). *Rev. Lecta* **21**, 47–54.

Cook, L.M. and Soltis, P.S. (2000) Mating systems of diploid and allotetraploid populations of Tragopogon (Asteraceae). II. Artificial populations. *Heredity (Edinb).* **84**, 410–415.

Costa, A., Guimarães-Dias, F. and Pérez-Maluf, R. (2008) Abelhas (Hymenoptera: Apoidea) visitantes das flores de urucum em Vitória da Conquista, BA. *Ciência Rural* **38**, 534–537.

Cuthbert, J.L. and McVetty, P.B.E. (2001) Plot-to-plot, row-to-row and plant-to-plant outcrossing studies in oilseed rape. *Can. J. Plant Sci.* **81**, 367–372.

Dequigiovanni, G., Ramos, S.L.F., Zucchi, M.I., Bajay, M.M., Pinheiro, J.B., Fabri, E.G., Bressan, E.A. and Veasey, E.A. (2014) Isolation and characterization of microsatellite loci for *Bixa orellana*, an important source of natural dyes. *Genet. Mol. Res.* **13**.

Devaux, C., Lepers, C. and Porcher, E. (2014) Constraints imposed by pollinator behaviour on the ecology and evolution of plant mating systems. *J. Evol. Biol.* **27**, 1413–1430.

Dick, C.W., Etchelecu, G. and Austerlitz, F. (2003) Pollen dispersal of tropical trees (*Dinizia excelsa*: Fabaceae) by native insects and African honeybees in pristine and fragmented Amazonian rainforest. *Mol. Ecol.* **12**, 753–764.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of Plant DNA from fresh tissue. *Focus (Madison).* **12**, 13–15.

Gaino, A.P.S.C., Silva, A.M., Moraes, M.A., Alves, P.F., Moraes, M.L.T., Freitas, M.L.M. and Sebbenn, A.M. (2010) Understanding the effects of isolation on seed and pollen flow, spatial genetic structure and effective population size of the dioecious tropical tree species *Myracrodruon urundeuva*. *Conserv. Genet.* **11**, 1631–1643.

Gomes, P., Alves, M., Draper, D., *et al.* (2009) floristic and vegetational aspects of an inselberg in the semi-arid region of northeast brazil. *Edinburgh J. Bot.* **66**, 329.

Hufford, K.M. and Hamrick, J.L. (2003) Viability selection at three early life stages of the tropical tree, *Platypodium elegans* (Fabaceae, Papilionoideae). *Evolution* **57**, 518–26.

Jarvis, A., Lane, A. and Hijmans, R.J. (2008) The effect of climate change on crop wild relatives. *Agric. Ecosyst. Environ.* **126**, 13–23.

Joseph, N. and Siril, E.A. (2014) Evaluation and selection of elite annatto (*Bixa orellana* L.) and identification of RAPD markers associated with yield traits. *Rev. Bras. Bot.* **37**, 1–8.

Joseph, N. and Siril, E.A. (2013) Floral Color Polymorphism and Reproductive Success in Annatto *(Bixa orellana* L.). *Trop. Plant Biol.* **6**, 217–227.

Joseph, N., Siril, E.A. and Nair, G.M. (2012a) Reproductive characterization and preliminary studies on controlled breeding of Annatto (*Bixa orellana* L.). *Plant Syst. Evol.* **298**, 239–250.

Joseph, N., Siril, E.A. and Nair, G.M. (2012b) Reproductive characterization and preliminary studies on controlled breeding of Annatto (*Bixa orellana* L.). *Plant Syst. Evol.* **298**, 239–250.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R. and Dhawan, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177**, 309–334.

Karasawa, M.M.G., Vencovsky, R., Silva, C.M., Zucchi, M.I., Oliveira, G.C.X. and Veasey, E.A. (2007) Mating system of Brazilian *Oryza glumaepatula* populations studied with microsatellite markers. *Ann. Bot.* **99**, 245–253.

Lombello, R.A. and Pinto-Maglio, C.A.F. (2014) Cytogenetics and Reproductive Biology of *Bixa orellana* L. (Bixaceae). *Cytologia (Tokyo).* **79**, 379–386.

Louette, D. (2000) Traditional management of seed and genetic diversity : what is a landrace ? In Genes in the field. On farm Conservation of crop diversity., pp. 109–142. IDRC, Ottawa, ON, CA.

Manoel, R.O., Alves, P.F., Dourado, C.L., Gaino, A.P.S.C., Freitas, M.L.M., Moraes, M.L.T. and Sebbenn, A.M. (2012) Contemporary pollen flow, mating patterns and effective population size inferred from paternity analysis in a small fragmented population of the Neotropical tree *Copaifera langsdorffii* Desf. (Leguminosae-Caesalpinioideae). *Conserv. Genet.* **13**, 613–623.

Meyer, R.S., DuVal, A.E. and Jensen, H.R. (2012) Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.* **196**, 29–48.

Moreira, P.A., Lins, J., Dequigiovanni, G., Veasey, E.A. and Clement, C.R. (2015) The Domestication of Annatto (*Bixa orellana*) from *Bixa urucurana* in Amazonia. *Econ. Bot.* **69**, 127–135.

Morton, J.F. (1960) Can Annato (*Bixa orellana L.*) an old source of food color, meet new needs for safe dye. *Florida State Hortic. Soc.*, 301–309.

Nazareno, A.G. and Reis, M.S. d. (2012) Linking Phenology to Mating System: Exploring the Reproductive Biology of the Threatened Palm Species *Butia eriospatha*. *J. Hered.* **103**, 842–852.

Nisar, N., Li, L., Lu, S., Khin, N.C. and Pogson, B.J. (2015) Carotenoid metabolism in plants. *Mol. Plant* **8**, 68–82.

Picanço-Rodrigues, D., Astolfi-Filho, S., Lemes, M.R., Gribel, R., Sebbenn, A.M. and Clement, C.R. (2015) Conservation implications of the mating system of the Pampa Hermosa landrace of peach palm analyzed with microsatellite markers. *Genet. Mol. Biol.* **38**, 59–66.

Ramos, S.L.F., Lopes, M.T.G., Lopes, R., Cunha, R.N.V. da, Macêdo, J.L.V. de, Contim, L.A.S., Clement, C.R., Rodrigues, D.P. and Bernardes, L.G. (2011) Determination of the mating system of Tucumã palm using microsatellite markers. *Crop Breed. Appl. Biotechnol.* **11**, 181–185.

Ritland, K. (1989) Correlated matings in the partial selfer Mimulus guttatus. *Evolution (N. Y).* **43**, 848–859.

Ritland, K. (2002) Extensions of models for the estimation of mating systems using n independent loci. *Heredity (Edinb).* **88**, 221–8.

Ritland, K. (2004) Multilocus mating system program MLTR. Version 3.1.

Ritland, K. and Jain, S. (1981) A Model for the Estimation of Outcrossing Rate and Gene Frequencies Using in Independent Loci. *Heredity (Edinb).* **47(1)**, 35–52.

Rivera-Madrid, R., Escobedo-GM, R.M.M., Balam-Galera, E., Vera-Ku, M. and Harries, H. (2006) Preliminary studies toward genetic improvement of annatto (*Bixa orellana* L.). *Sci. Hortic. (Amsterdam).* **109**, 165–172.

Rivera, D.I. and Flores, E.M. (1988) Morfología floral del achiote, *Bixa orellana* L. (Bixaceae). *Rev. Biol. Trop. J. Trop. Biol. Conserv.* **36**, 499–509.

Ross-Ibarra, J., Morrell, P.L. and Gaut, B.S. (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci. U. S. A.*, 8641–8.

Sandy-Cuen, P.M. and Becerra, R. (2003) Manejo campesino de recursos naturales. El achiote. *BioDiversitas* **7**, 7–11.

Sebbenn, A.M., Kageyama, P.Y., Siqueira, A. and Zanatto, A. (2000) Taxa de cruzamento em populações de *Cariniana legalis* (Mart.) O. Ktze.: Implicações para a conservação e o melhoramento genético. *Sci. Florestalis* **58**, 25–50.

Slatkin, M. (1987) Gene flow and the geographic structure of natural populations. *Science* **236**, 787–92.

Valdez-Ojeda, R., Quiros, C.F., Lourdes Aguilar-Espinosa, M. de, Rivera-Madrid, R., Aguilar-Espinosa, M. de L. and Rivera-Madrid, R. (2010) Outcrossing rates in annatto determined by sequence-related amplified polymorphism. *Agron. J.* **102**, 1340–1345.

Vallejo, C. (1991) Evaluacion de germoplasma de achiote *Bixa orellana* L.: Estudios básicos sobre asociaciones fenotípicas y biologıa floral. *Acta Agron.*, 7–20.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. (2016) Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **39**, 312–328.

Ward, M., Dick, C.W., Gribel, R. and Lowe, A.J. (2005) To self, or not to self... a review of outcrossing and pollen-mediated gene flow in neotropical trees. *Heredity (Edinb).* **95**, 246–254.

Yolmeh, M., Habibi Najafi, M.B. and Farhoosh, R. (2014) Optimisation of ultrasound-assisted extraction of natural pigment from annatto seeds by response surface methodology (RSM). *Food Chem.* **155**, 319–324.

7. DOMESTICATION OF ANNATTO FROM *BIXA ORELLANA* VAR. *URUCURANA*: EVIDENCES BASED ON SSR MARKERS AND CHLOROPLAST DNA

Abstract

Plant domestication is an evolutionary process that leads to several modifications in plants to increase adaptation to cultivation and utilization by humans. These modifications may decrease the fitness of plants in the wild habitat but increase it for human exploitation. Amazonia is an important center of biodiversity and many species were domesticated in this area. Annatto (*Bixa orellana* L.) is a shrubby plant domesticated in Amazonia from wild annatto (*B. orellana* var. *urucurana*) populations. In this study, we used molecular markers based on microsatellites and cpDNA, to compare the levels of genetic diversity and genetic structure between wild and cultivated annatto, to identify the existence of hybridization between wild and cultivated annatto, as well as providing good evidence of the likely origin of domestication of annatto in Brazilian Amazonia. We evaluated 490 cultivated and 170 wild annatto samples from Brazilian Amazonia, using 16 microsatellite loci and two cpDNA regions. We found a clear separation between the two varieties, cultivated and wild annatto, identified in the Bayesian, Neighbor-Joining trees and PCoAs analyses. High values of F_{ST}, both in the SSR and cpDNA analyses, support the separation into two groups. Wild accessions presented higher rates of diversity in relation to cultivated annatto, partly because these populations do not suffer anthropic selection, as in the cultivated varieties. Our data suggest the existence of genetic relationship between wild and cultivated annatto, indicated by moderate levels of gene flow between populations. Our results highlight the proximity between groups of cultivated and wild accessions from Rondônia and the Madeira River basin. This proximity provides indications that annatto started its domestication in this area from *B. orellana* var. *urucurana*.

Keywords: Brazilian Amazonia; Domestication; Annatto; CpDNA; SSR; Wild

7.1. Introduction

Plant domestication is an evolutionary process that leads to increased adaptation of plants to cultivation and utilization by humans (Ladizinsky, 1985; Brown, 2010). It is a continuing process over a long period of time, leading to significant changes in genotype and phenotype of the plants that are subjected to this process (Gepts, 2004). In this process wild populations are subjected to selection and as a consequence modifications occur which make them dependent on man for reproduction and survival (Ladizinsky, 1985). These modifications may decrease the fitness of plants in wild habitat but increase it for human exploitation (Meyer and Purugganan, 2013). Domesticated plants present a number of morphological changes when compared to their wild ancestors, known as the domestication syndrome (Harlan, 1992; Gepts, 2004), and have undergone so many changes that in some cases it is difficult to establish the relationships between them and their wild ancestor

(Ladizinsky, 1985). This distinction between wild and cultivated plants is one of the basic questions of plant domestication (Terrell *et al.*, 2003).

Amazonia is the world's most important center of biodiversity and many species were domesticated in this area, including cassava, cacao, sweet potato, pineapple and numerous fruit trees and palms (Clement, 1989; Clement *et al.*, 2010; 2016). Annatto (*Bixa orellana* L.) is a shrubby plant domesticated in Amazonia (Sandy-Cuen and Becerra, 2003; Clement *et al.*, 2010; Moreira *et al.*, 2015; Clement *et al.*, 2016). Historical evidences indicate its extensive distribution and cultivation in American tropics and subsequently its spread worldwide after the European contact (Leal and Clavijo, 2010; Akshatha *et al.*, 2011). Annatto was used in many different cultural contexts because of the red–orange pigment present in the arils around the seeds (León, 2000).

Moreira *et al.* (2015) proposed that *B. urucurana* is the wild ancestor of cultivated annatto, *B. orellana*. They also accepted Pilger's proposal, published by Kuntz, that *urucurana* is a variety of *B. orellana* (*B. orellana* var. *urucurana* (Willd.) Kuntze ex Pilg.). *Bixa orellana* var. *urucurana* occurs in open forests or anthropogenic landscapes in riparian environments. Cultivated annatto (*B. orellana*) produces large amounts of pigment around its seeds, while the variety *urucurana* presents variable amounts. Gene flow between the varieties results in variability in pigment production (Clement *et al.*, 2016). Cultivated annatto also produces larger fruit sizes and number of seeds per fruit, as a result of the domestication syndrome for this species. But a surprising difference between the two varieties is that cultivated annatto is dehiscent, while the wild annatto variety is indehiscent, remaining closed after maturation (Moreira *et al.*, 2015).

The exact location where annatto was domesticated is still unclear. The only archaeological record of annatto in Amazonia is dated to 2400 years before present in southwestern Amazonia, in Llanos de Mojos (Bolivia) (Erickson, 1995). The upper Madeira River basin, in the periphery of Amazonia, has also been recognized as a probable region of the crop's origin (Piperno and Pearsall, 1998) .

Microsatellite or SSR (*Simple Sequence Repeat*) markers are important tools to access genetic diversity and genetic structure of populations. Microsatellite markers feature hypervariability and co-dominance, and besides presenting highly reproducible results, it is widely distributed in the nuclear genome (Kalia *et al.*, 2011; Vieira *et al.*, 2016). Chloroplast DNA (cpDNA) sequences have added a great understanding to the complicated evolutionary histories of plants. This marker has uniparental inheritance and does not suffer recombination at the meiosis process, as it occurs in nuclear DNA, being useful in studies of dispersion of

crop species (Clement *et al.*, 2010). Additionally, many studies have revealed intraspecific and interspecific sharing of cpDNA haplotypes, highlighting the potential for revealing hybridization and introgression among closely related taxa (Jakob and Blattner, 2006; Yano *et al.*, 2010).

In this study we used microsatellite markers and cpDNA variations to answer the following questions: a) what are the levels of genetic diversity and genetic structure among cultivated annatto, compared to wild annatto?; b) is there any evidence of gene flow, and therefore, of hybridization between wild and cultivated annatto?; c) by the evidences gathered in this study, is it possible to suggest the most likely origin of domestication of annatto in Brazilian Amazonia?

7.2. Material and methods

7.2.1. Field sampling

A total of 630 samples of annatto from Brazilian Amazonia was analyzed in this study (Table 7.1; Figure 7.1). From these, 170 samples were identified as wild annatto (*Bixa orellana* var *urucurana*) and 460 as cultivated annatto (*Bixa orellana*). The domesticated varieties were obtained in homegardens of riverside communities in Amazonia along Madeira River, Branco River, Negro River, Solimões River and Amazonas River, in the states of Rondônia, Roraima, Amazonas and Pará. We also sampled annatto varieties from farmer´s fields in the states of Rondônia and Pará. The wild samples were obtained from 10 municipalities in the states of Rondônia, Pará and Roraima, also in Brazilian Amazonia: Cabixi, RO; Cerejeiras, RO; Corumbiara, RO; São Francisco do Guaporé, RO; Jí-Paraná, RO; Ariquemes, RO; Monte Alegre, PA; Almeirin, PA; Bom Jesus do Tocantins, PA; e Mucajaí, RR. Geographical position of each sampled site was registered using the Global Positioning System (GPS). We collected leaves for genetic analyses of each plant (N=630). DNA was extracted from leaves using the CTAB 3 % protocol (Doyle and Doyle, 1990) with minor modifications. DNA concentration was determined using 1 % agarose gels stained with GelRed (Biotium), comparing with 25, 50 and 100 ng/µL concentrations of lambda phage DNA. We used uniform DNA concentrations of 5 ng/µL for the Polymerase Chain Reactions.

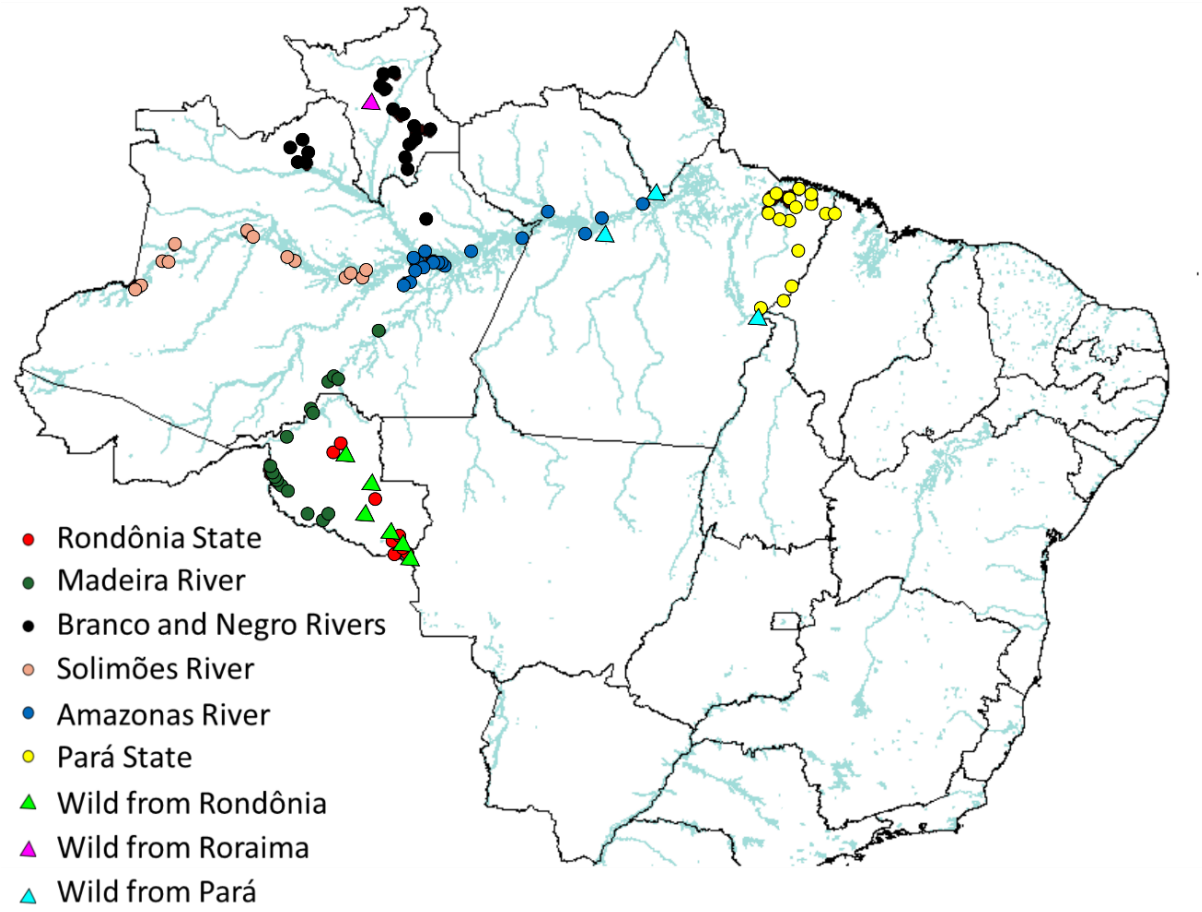

Figure 7.1. Geographic distribution of collection localities of cultivated (*Bixa orellana*) and wild (*B. orellana* var. *urucurana*) annatto samples in Brazilian Amazônia

7.2.2. Microsatellite data

For the microsatellite data, all the 630 wild and domesticated samples of annatto from Brazilian Amazonia were analyzed with 16 microsatellite loci. The genotyping methodologies are described in Chapters 2 and 5. We estimated the following parameters for genetic diversity: mean number of alleles per locus (*A*), mean number of effective alleles per locus (N_E), observed (H_O) and expected heterozygosity (H_E). The genetic diversity indices were estimated using *Adegenet* (Jombart and Ahmed, 2011) for R (R Core Team, 2015).

The population structure was hypothesized using a Bayesian analysis with *Structure* software (Pritchard *et al.*, 2000). The number of clusters (*K*) was estimated by performing ten independent runs for each *K* varying from 1 to 10, using 500,000 MCMC repetitions with a burn-in period of 200,000. The model with correlated allele frequencies and admixture was carried out. The most likely number of clusters was evaluated with the ad hoc method of Evanno *et al.* (2005).

Population structure levels were also computed using a molecular variance analysis (AMOVA) in the software Arlequin version 3.5.1.2 (Excoffier and Lischer, 2010). Pairwise genetic differentiation among groups were estimated by Weir and Cockerham's (1984) F_{ST} with significance based upon 9,999 bootstraps replicates performed with Arlequin version 3.5.1.2 (Excoffier and Lischer, 2010).

We constructed a Neighbor-Joining (Saitou and Nei, 1987) dendrogram to represent the relationships between groups. Cavalli-Sforza and Edwards' (Cavalli-Sforza and Edwards, 1967) chord distance (D_{CE}) was obtained with *MSA 4.05* (Dieringer and Schlötterer, 2003), and the dendrogram was built in *Phylip 3.5* (Felsenstein, 2005), and formatted in *FigTree 1.4.1* (http://tree.bio.ed.ac.uk/software/figtree/). Genetic diversity of annatto groups and individuals was explored with Principal Coordinates Analysis (PCoA) executed with *GenAlEx 6.5* (Peakall and Smouse, 2012). In the PCoA and NJ analyzes the groups of wild samples from Rondônia, Roraima and Pará were dismembered in small sub-groups according to their municipality of origin.

7.2.3. Chloroplast DNA data

From all the available samples used for genotyping with microsatellite markers, we selected 240 representative accessions of all the areas studied for sequencing of cpDNA regions. Initially, different regions of chloroplast DNA were submitted to amplification to identify those that presented better results in terms of sequencing, as well as polymorphisms. The following 14 cpDNA regions were tested in a group of 10 individuals: *rpl32-trnL, atpI-atpH, psbJ-petA, psbA-trnH, psbD-trnT, atpBrbcL, rpl16-intron, rpl32-ndhF, rpoB-trnC, rps16-trnk, rps16-trnQ, trnS-trnG, trnV-ndhC* and *accD-psaI* (Shaw *et al.*, 2005; Shaw *et al.*, 2007). The amplification reactions of the cpDNA regions were performed according to Shaw *et al.* (2005). Fragments were purified and sequenced with the Big Dye 3.1 kit (Applied Biosystems, Foster City, CA, USA), with standard protocol for all reactions by the company Macrogen Inc.

The cpDNA sequences obtained for each individual were edited in the PreGap4 and Gap4 software (Staden *et al.*, 1998). Alignment of the sequences was performed with the Muscle algorithm (Edgar, 2004) powered in the MEGA 7 software (Kumar *et al.*, 2016). Polymorphism levels were calculated by the DNAsp software (Librado and Rozas, 2009). The following genetic diversity parameters were estimated: number of haplotypes (Nh), number of

polymorphic sites (S), haplotype diversity (Hd), and nucleotide diversity (π). We tested for deviation of the standard neutral model using the Tajima's D and the Fu and Li D tests (Fu and Li, 1993; Tajima, 1989). Population structure levels (AMOVA) and pairwise genetic differentiation among groups were estimated by Weir and Cockerham's (1984). F_{ST} with significance based upon 9,999 bootstraps replicates was obtained with Arlequin version 3.5.1.2 (Excoffier and Lischer, 2010). We also built haplotype networks with the median-joining method, with Network 4.6.1 software (Bandelt *et al.*, 1999).

7.3. Results

7.3.1. Microsatellite analysis

Genetic diversity estimates showed a mean number of alleles per locus ($\bar{A}$) of 5.813, with a higher average for the wild samples (6.021) when compared to cultivated annatto (5.708) (Table 7.1). The wild samples from Rondônia showed the highest mean number of alleles per locus (9.563) and mean number of effective alleles per locus (4.370). The mean values of observed (H_O) and expected (H_E) heterozygosities for all groups were 0.205 and 0.532, respectively, with lower H_O values observed for both wild and cultivated annatto. Wild annatto groups presented the highest values for both parameters (Table 7.1).

Table 7.1 Genetic parameters estimated for six groups of cultivate annatto and three of wild annatto, totaling nine groups including sample size (N) mean number of alleles per locus ($\bar{A}$), mean number of effective alleles per locus (N_E), observed (H_O) and expected (H_E) heterozygosity

Group	N	$\bar{A}$	N_E	H_O	H_E
Rondônia State	126	6.188	2.229	0.158	0.452
Madeira River	66	5.063	2.358	0.126	0.491
Branco and Negro Rivers	53	6.750	2.340	0.139	0.496
Solimões River	39	5.188	2.512	0.105	0.506
Amazonas River	40	4.313	2.193	0.093	0.432
Pará State	136	6.750	2.502	0.129	0.500
All cultivated annatto	460	5.708	2.356	0.125	0.479
Wild from Rondônia	149	9.563	4.370	0.455	0.670
Wild from Roraima	9	4.313	3.153	0.223	0.604
Wild from Pará	12	4.188	3.176	0.415	0.635
All wild annatto	170	6.021	3.566	0.364	0.636
Overall	630	5.813	2.759	0.205	0.532

To investigate the genetic relationships among samples and to search for evidences of genetic admixture between wild and cultivated annatto genotypes, we performed Bayesian analysis with the *Structure* software. The 630 annatto plants were grouped into genetic clusters by the *Structure* simulations, and two clusters (K = 2) were identified as the most likely structure (Figure 7.2) based on the ad hoc ΔK approach (Evanno *et al.*, 2005). The two clusters correspond to cultivated and wild annatto (Figure 7.3).

In addition, we chose to evaluate the structuring using a larger number of groups (Figure 7.3), varying the number of groups (K) in 2, 3, 4 and 10. The wild samples were more structured and remained with the same structure even when advancing until K = 10. The same happened with the group of samples from East Pará, that are quite distanced from the other samples. While wild annatto samples were maintained as a structured group, cultivated samples were classified into two and three groups for K=3 and K=4, respectively, sharing more genetic information between groups, except for the East Pará group. With K=3, the two groups among the cultivated samples were a first group with samples from Rondônia State, Madeira River and Pará State, while the second group classified the samples from Branco, Solimões and Amazonas Rivers, and part of Pará State. However, with K=4, the Pará State samples included in the first group with K=2, among the cultivated samples, were now classified in a third isolated group.

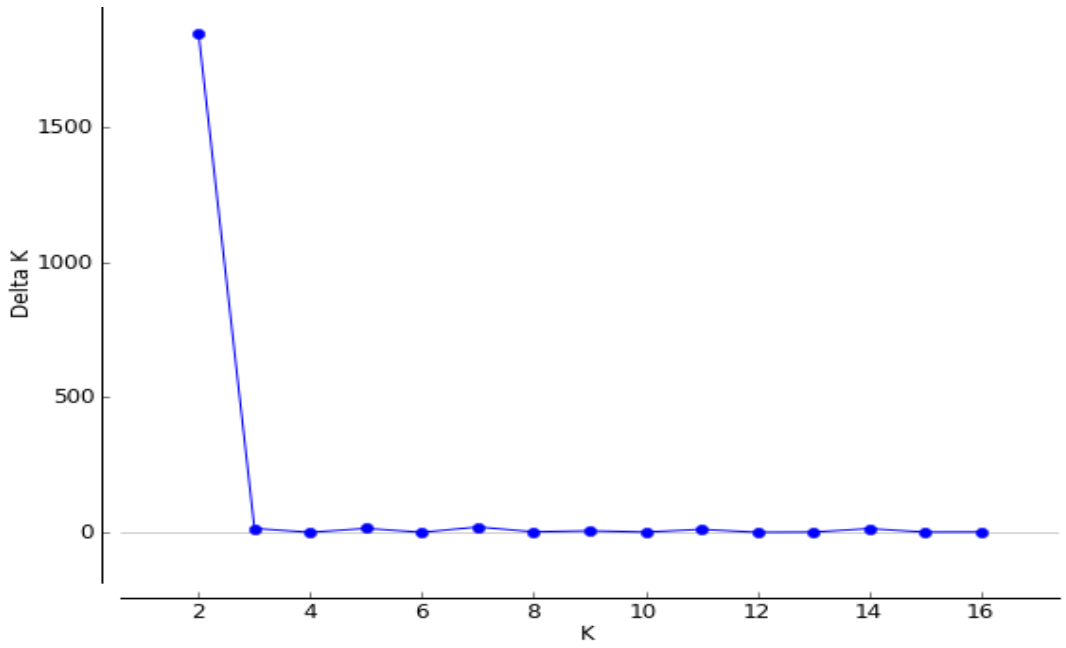

Figure 7.2. Evanno *et al.* (2005) plot detecting the number of *K* groups that best fit the data for 16 SSR markers and 630 annatto accessions

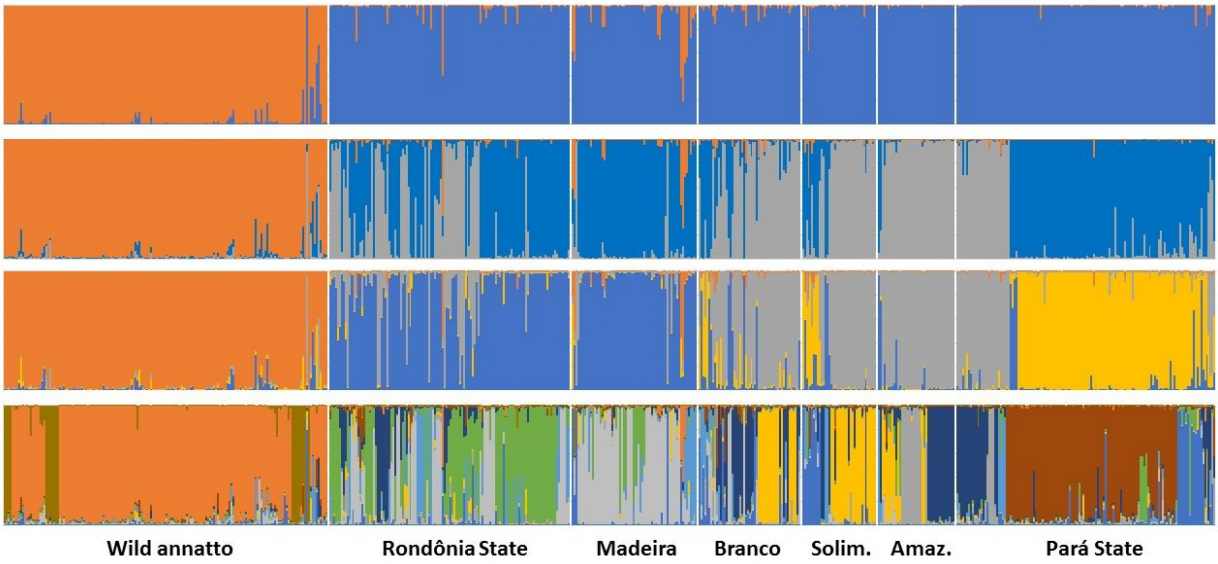

Figure 7.3. Graphical display of the Bayesian analyses based on 16 SSR loci and accessions of *Bixa orellana* and *B. orellana* var. *urucurana* populations with *K* = 2, 3, 4 inferred with the Structure software. Madeira = Madeira River; Branco = Branco and Negro Rivers; Solim. = Solimões River; Amaz. = Amazonas River

The projection of the 630 individuals on a two-dimensional plane defined by the first two principal coordinates (Figure 7.4), which explained 52% of total variation, clustered almost all wild annatto samples in the top right area of the plot. The first coordinate separated the wild samples together with the cultivated samples from Rondônia and Madeira River in the left area. The analysis also showed that cultivated samples from the State of Pará are located in the top left area of the plot, and are genetically more distant from all wild and cultivated samples, although some of these samples are closer to those from Rivers Amazonas, Branco and Solimões, as shown in the Bayesian analysis (Figure 7.3). It is also possible to identify that annatto cultivated along riversides and traditional farmers from the Madeira River basin and also some samples from other parts of the state of Rondônia, are closer to the wild samples (Figure 7.4).

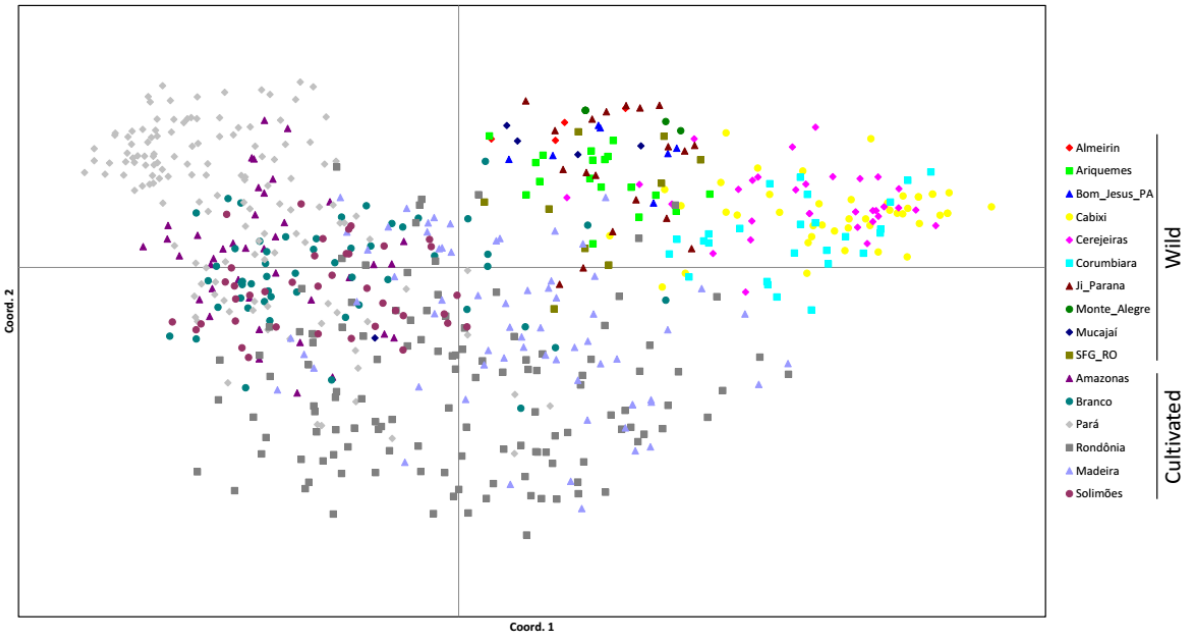

Figure 7.4. Principal coordinate analysis (PCoA) plot based on Cavalli-Sforza and Edwards (1967) Chord distance among 460 cultivated annatto (*Bixa orellana*) and 170 wild annatto (*B. orellana* var. *urucurana*) with 16 SSR loci. The groups of wild accessions from Rondônia, Roraima and Pará were dismembered in small populations according to the municipality of origin of the samples

The evaluation of the relationships among the annatto groups using PCoA (Figure 7.5), which explained 60% of total variation, showed a tendency to group the samples according to the type of variety (wild and domesticated) and their geographical origin. The first coordinate separated the cultivated types (on the left) from the wild types (on the right). The second coordinate separated the groups from the North (Top) from the Southern groups (bottom). The PCoA also shows that the cultivated annatto groups from the Madeira River basin and the state of Rondônia are closer to the wild groups from the same area, with both types of groups allocated in the bottom of half of the second coordinate.

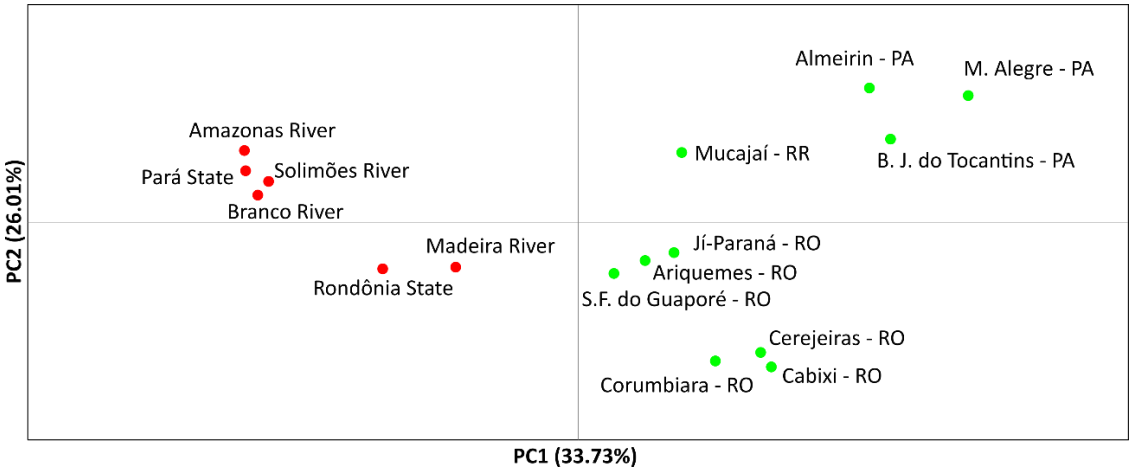

Figure 7.5. Principal coordinate analysis (PCoA) based on Cavalli-Sforza and Edwards (1967) Chord distance among groups of *Bixa orellana* (red circles) and *B. orellana* var. *urucurana* (green circles) with 16 SSR loci. The groups of wild accessions from Rondônia, Roraima and Pará were dismembered in small populations according to the municipality of origin of the samples

The dendrogram (Figure 7.6) showed the same pattern observed in the Bayesian and dispersion analyzes (PCoA). NJ dendrogram also highlighted the proximity between groups of accessions from Rondônia and the Madeira River basin. This pattern is observed in all analyses.

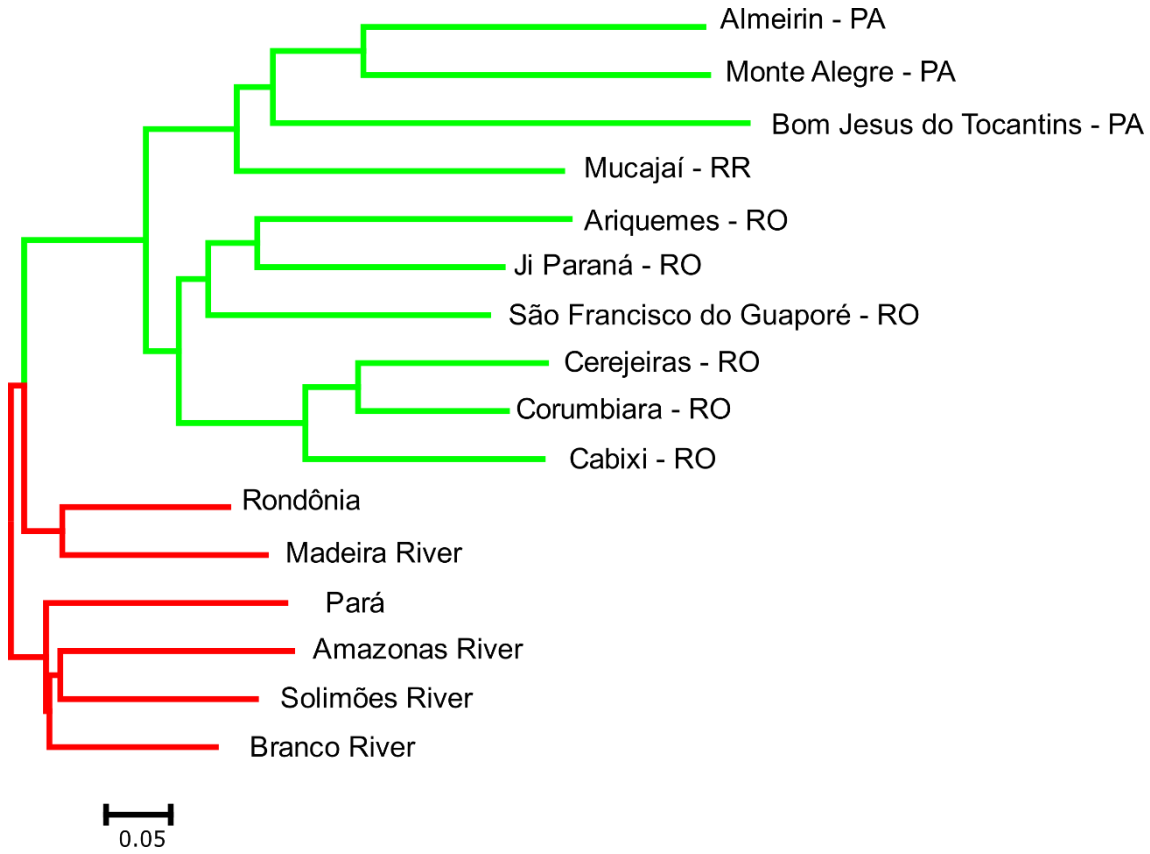

Figure 7.6. Neighbor-joining dendrogram for *Bixa orellana* var. *urucurana* (green) and *B. orellana* (red) groups based on Cavalli-Sforza and Edwards (1967) Chord distance estimated with 16 nuclear microsatellites (SSR). The groups of wild accessions from Rondônia, Roraima and Pará were dismembered in small populations according to the municipality of origin of the samples

AMOVA was performed considering all cultivated samples as one group (cultivated) and the wild samples as another group (wild). And another analysis was made considering the samples divided into nine groups (six cultivated plus three wild groups, Table 7.1). Cultivated and wild samples were significantly divergent (F_{ST} = 0.158) (Table 7.2). However, the highest divergence was observed among the nine groups (F_{ST} = 0.182). AMOVA also revealed that the highest proportion of variation was found within wild and cultivated samples (84.1%) and within the nine groups (81.8%).

Table 7.2. Molecular variance analysis (AMOVA) performed for SSR and cpDNA for different groups of wild (*Bixa orellana var. urucurana*) and cultivated (*B. orellana*) annatto

Source of variation	SSR		cpDNA	
	Percentage of variation	F-statistics	Percentage of variation	F-statistics
Between wild and cultivated samples	15.84	$F_{ST} = 0.158^{*}$	11.93	$F_{ST} = 0.119^{*}$
Within wild and cultivated samples	84.16		88.07	
Between groups	18.16	$F_{ST} = 0.182^{*}$	43.24	$F_{ST} = 0.432^{*}$
Within groups	81.84		56.76	

*significant at $p < 0.05$

Pairwise F_{ST} estimates (Table 7.3) also show high divergence between cultivated and wild varieties, as well as different levels of divergence among groups from major Amazonia rivers and the adjacent regions. The highest divergences among cultivated samples were between samples from Rondônia (Rondônia and Madeira River) and Amazonas River, and also from Pará and Rondônia. Among the wild samples, the greatest divergences were among wild Rondônia and wild Pará, followed by wild Rondônia and wild Roraima groups.

Table 7.3. Pairwise F_{ST} (Weir & Cockerham, 1984) estimates among groups of cultivated and wild annatto based on the variation of 16 SSR. Boldfaced values are significant based upon 9,999 bootstrap replicates

Group	RO State	Madeira	Branco	Solimões	Amazonas	PA State	Wild_RO	Wild_PA
RO State								
Madeira	0.077							
Branco	**0.114**	**0.096**						
Solimões	**0.117**	**0.126**	**0.037**					
Amazonas	**0.168**	**0.192**	**0.084**	**0.082**				
PA State	**0.175**	**0.181**	**0.087**	**0.107**	**0.131**			
Wild_RO	**0.180**	**0.150**	**0.191**	**0.201**	**0.247**	**0.244**		
Wild_RR	**0.269**	**0.229**	**0.217**	**0.210**	**0.249**	**0.248**	**0.172**	
Wild_Pa	**0.325**	**0.289**	**0.284**	**0.274**	**0.315**	**0.323**	**0.194**	**0.109**

7.3.2. Chloroplast DNA analysis

In the cpDNA analysis, from the amplification and sequencing test performed with 10 individuals, we identified that the *rpl32-trnL* and *accD-psaI* regions presented the best results and were, therefore, selected for amplification in the 240 samples. These were submitted to sequencing of the two cpDNA regions and after analysis, 172 samples presented sequences with high quality for the two regions analyzed.

Sequenced regions of *rpl32-trnL* and *accD-psaI* cpDNA resulted in a total alignment of 1487 bases. The *rpl32-trn*l region exhibited 745 bases while the *accD-psaI* region showed 742 bases. Diversity indexes were calculated for the groups sampled, and the number of

haplotypes ranged from two in cultivated groups from Amazonas River and Pará to nine in the wild group from Rondônia, respectively. The number of polymorphic sites also varied from 2 to 34 in cultivated group from Amazonas River and wild group from Rondônia, respectively. Both haplotype and nucleotide diversity indices were higher for cultivated than wild groups. The haplotype diversity varied from 0.282 for the group of cultivated annatto from Pará to 0.750 for the group of wild samples from Pará and for cultivated annatto from Branco River. The nucleotide diversity varied from 0.0004 for the wild group from Roraima to 0.0059 cultivated group from Rondônia. The Tajima's D and the Fu and Li tests results did not significantly deviate from the standard neutral model (Table 7.4).

Table 7.4 - Analysis of genetic diversity with cpDNA for the sampled populations of *Bixa orellana* and *B. orellana* var. *urucurana*. *N* = number of individuals, *Nh* = number of haplotypes, *S* = number of polymorphic sites; *Hd* = haplotype diversity; π = nucleotide diversity; *D* = Tajima'a D test; *Fs* = Fu's Fs test

Groups	***N***	***Nh***	***S***	***Hd***	***π***	***D***	**P-value**	***Fs***	**P-value**
Cultivated	96	11	19	0.704	0.0039	-0.390	0.408	3.095	0.856
Wild	76	13	37	0.589	0.0025	-1.035	0.145	-0.840	0.414
RO State[1]	13	3	15	0.692	0.0059	0.951	0.884	8.375	0.996
Madeira	18	4	3	0.608	0.0008	0.001	0.598	0.062	0.511
Branco	17	7	18	0.750	0.0023	-0.438	0.376	-0.346	0.436
Solimoes	22	4	4	0.619	0.0009	-0.328	0.365	0.476	0.595
Amazonas	13	2	2	0.385	0.0006	0.426	0.807	1.907	0.789
PA State	13	2	14	0.282	0.0030	-0.274	0.301	7.439	0.997
W_RO	64	9	34	0.533	0.0025	-0.991	0.170	1.338	0.759
W_RR	4	2	1	0.500	0.0004	-0.612	0.385	0.172	0.346
W_PA	8	4	4	0.750	0.0013	0.283	0.625	-0.240	0.313

[1] RO State = Cultivated samples from Rondônia; Madeira = Madeira River; Branco = Branco and Negro Rivers; Solimões = Solimões River; Amazonas = Amazonas River; PA State = samples from Pará; W_RO = Wild samples from Rondônia; W_RR = Wild samples from Roraima; W_PA = Wild samples from Pará.

Chloroplast differentiation according to AMOVA (Table 7.2) was significant between cultivated and wild samples (F_{ST} = 0.119), similar to the SSR data. However, the highest divergence observed was between groups (F_{ST} = 0.432), considering the nine groups of cultivated and wild samples. The highest proportion of variation was found within wild and cultivated groups (81.1%). Considering the nine groups, the percentage within groups (56.8%) was also higher than among groups with cpDNA, but lower than with SSR markers (81.8%).

Pairwise F_{ST} estimates (Table 7.5) showed high divergence between cultivated groups from Pará and Rondônia from major Amazonia rivers, including Madeira River. The wild group from Pará also showed high divergence in relation to the other groups.

Table 7.5. Pairwise F_{ST} (Weir & Cockerham, 1984) estimates among groups of cultivated (*Bixa orellana*) and wild (*B. orellana* var. *urucurana*) annatto based on chloroplast genetic variation. Boldfaced values are significant based upon 9,999 bootstrap replicates

Group	RO State	Madeira	Branco	Solimões	Amazonas	PA State	Wild_RO	Wild_RR
RO State								
Madeira	**0.407**							
Branco	**0.269**	-0.003						
Solimões	**0.434**	**0.119**	**0.095**					
Amazonas	**0.380**	-0.026	-0.0098	**0.256**				
PA State	**0.207**	**0.808**	**0.703**	**0.815**	**0.802**			
Wild_RO	**0.382**	0.054	0.037	**0.191**	0.010	**0.718**		
Wild_RR	**0.266**	**0.149**	-0.001	**0.435**	0.067	**0.745**	-0.063	
Wild_PA	**0.434**	**0.572**	**0.384**	**0.644**	**0.583**	**0.783**	**0.312**	**0.509**

The analyses of the two chloroplast regions, *rpl32-trnL* and *accD-psaI*, resulted in 22 haplotypes (Table 7.4 and Figure 7.7). The most frequent haplotype was H1 occurring in 90 individuals and in all populations, both wild and cultivated types. The second most frequent haplotype was H2 occurring in 27 cultivated individuals from the major Rivers of Amazonia. The third most frequent haplotype was H5, occurring in 18 cultivated individuals from Pará and Rondônia, and an individual from Branco River. The distribution of haplotypes in the haplotype network revealed a close genetic relation between *B. orellana* and *B. orellana* var. *urucurana*. Both varieties share a large number of haplotypes. Wild samples are distributed in 13 out of 22 haplotypes, while the cultivated samples are distributed in 11 haplotypes.

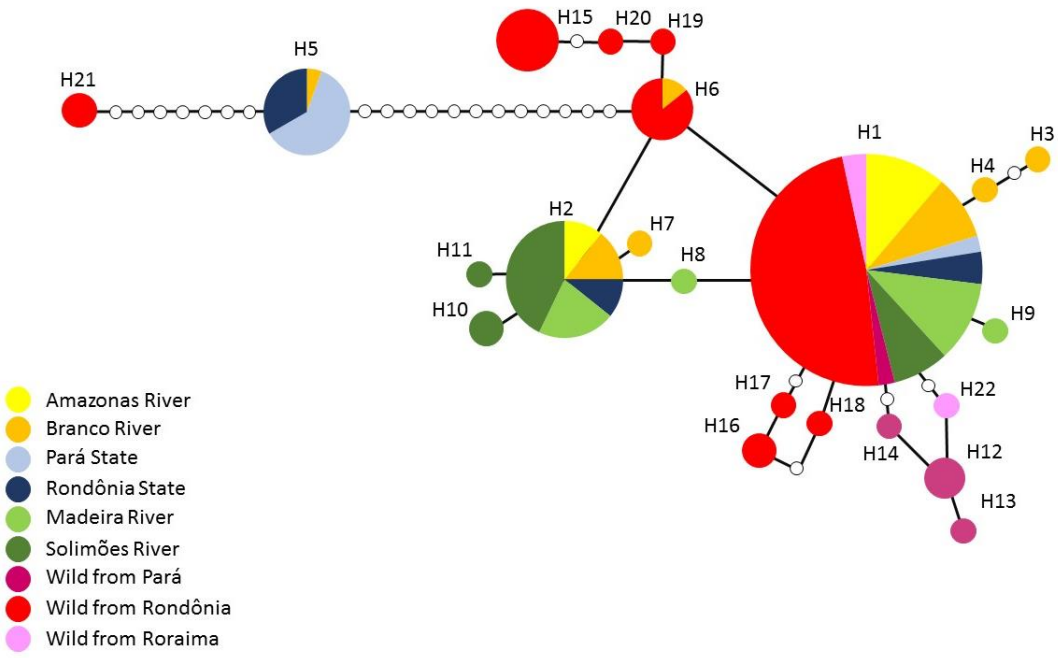

Figure 7.7. Median-joining network of chloroplast genetic variation for wild and cultivated annatto along major Amazonian rivers and adjacent regions in Brazilian Amazonia. Each circle represents a distinct haplotype, and size of circles is proportional to their frequencies

7.4. Discussion

Most domesticated crops generally show low genetic diversity compared to their wild relatives (Ladizinsky, 1985) as they undergo different selection pressures on productivity and yield (Gepts, 2004). In fact, levels of genetic diversity of wild annatto groups revealed by SSR analysis is substantially higher than that of the cultivated samples, as measured by observed (H_O) (reduction of 65 %) and expected heterozygosity (H_E) (reduction of 25%), and number of alleles per loci (reduction of 6 %) (Table 7.1). Gepts (2004) also highlighted that a common feature among most of the domesticated plants is a marked genetic bottleneck. The same pattern was observed in many other studies such as pigeonpea (*Cajanus cajan*) (Kassa *et al.*, 2012; Yang *et al.*, 2006), tomato (*Solanum lycopersicum*) (San-San-Yi *et al.*, 2008), coffee (*Coffea arabica*) (López-Gartner *et al.*, 2009), beans (*Phaseolus vulgaris*) (Bitocchi *et al.*, 2013), peach (*Prunus persica*) (Cao *et al.*, 2014), carrots (*Daucus carota*) (Rong *et al.*, 2014) and lima beans (*P. lunatus*) (Andueza-Noh *et al.*, 2015).

Different forces can cause reduction in genetic diversity of cultivated populations. Selection and dispersal of small groups of individuals from their wild habitats to new anthropogenic areas occur during the initial formation of a domesticated lineage ("the founder effect"), and directional selection occurs for genes associated with domestication traits, because people select individuals with more desirable morphological traits for future cultivation (Zohary, 2004; Kassa *et al.*, 2012). Intensive breeding can also cause further strong reductions to genetic diversity (Tanksley, 1997). The impact of these processes on genetic diversity are well described for major crops such as wheat, barley and maize, however, for many minor crops, the circumstances of domestication are poorly described (Matsuoka *et al.*, 2002; Dubcovsky and Dvorak, 2007; Morrell and Clegg, 2007).

Structure simulations identified two clusters among all samples evaluated, corresponding to wild and domesticated annatto. Besides the clear distinction, the existence of a small amount of admixture among wild and cultivated genomes was identified. Many crops are still growing alongside their wild relatives, increasing the opportunity for gene flow between cultivated and wild populations (Kassa *et al.*, 2012). Traditional farmers report that there are crosses between the two varieties of annatto, therefore, they tend to eliminate wild populations to avoid possible production losses (Moreira *et al.*, 2015). The admixture between the two types also suggest the possibility of hybridization between the two varieties,

confirming the hypothesis of Moreira *et al.* (2015) that wild annatto is a variety of *Bixa orellana.*

When the number of groups (*K*) increased to 4 and 10 in *Structure* analysis, wild annatto samples maintained structuring, while cultivated samples showed further levels of structure among the samples, with three well defined groups at K=4, although there was a sharing of genetic information between groups. This can be due to exchange of seeds among farmers and because some farmers cultivate several varieties in the same field. It is also common for farmers, when moving to another place, to carry seeds with them to start a new plantation. The influence of man on the dispersion of annatto seeds is well evidenced in this study.

NJ dendrogram and PCoAs highlight the proximity between groups of cultivated and wild accessions from Rondônia and the Madeira River basin. This proximity provides indications that annatto has started its domestication in this area from *B. orellana* var. *urucurana*.

The oldest archaeological record of annatto was found in PlumPiece, a pre–Arawak site dated to 3700–3600 before present on the Caribbean island of Saba (Moreira *et al.*, 2015). The geographical distribution of *B. orellana* var. *urucurana* does not include the Caribbean, and because of that this area was excluded as a candidate to origin of annatto domestication. The only archaeological record of annatto in Amazonia comes from the Llanos de Mojos, in Bolivia, and is dated to 2400 years before present (Erickson, 1995). In Amazonia, plant domestication started earlier than 8000 before present. Mostly in the periphery where wild populations of some domesticated plants have been identified using genetic and morphological analyses (Clement *et al.*, 2010; Levis *et al.*, 2017). Piperno and Pearsall (1998) also highlighted the importance of the periphery, in the extreme northwestern Amazonia and the adjacent Llanos of the Orinoco River basin, as well as in southwestern Amazonia, especially the Llanos de Mojos, in Bolivia. The upper Madeira River Basin drains a large portion of southwestern Amazonia and has also been recognized as a probable region of crop origins (Clement *et al.*, 2016; Piperno and Pearsall, 1998). This area presents an ecological complexity that offered numerous useful plants to hunter-gatherers, and some of these plants were domesticated in different areas of the upper Madeira basin. The upper Madeira basin was the home of numerous complex societies. It is the homeland of Tupi groups and it is adjacent to the probable origin of Arawak, the two most expansive language families of South America (Almeida and Neves, 2015; Santos *et al.*, 2015; Clement *et al.*, 2016).

According to Levis *et al.* (2017), plant species that responded well to selection and propagation were widely cultivated and dispersed within and outside their natural range. Historical evidences indicate that annatto extensive distribution and cultivation began initially in the American tropics and nowadays has been cultivated in the rest of the world, suggesting a well adaptation for different ecoclimatic conditions (Akshatha *et al.*, 2011; Leal and Clavijo, 2010).

Our chloroplast analysis suggests a single primary lineage, surrounded by several lower-frequency haplotypes. Therefore, these central high-frequency haplotypes, highly represented by wild samples from Rondônia, represent the ancestral haplotypes, with the low-frequency haplotypes more recently derived. This result conflicts with the microsatellite findings (*Structure*, *K*=2) which displayed a high discrimination between the two varieties of annatto. These differences could be due to the fact that unlike microsatellites, cpDNA is not affected by recombination and present low-rate of mutations (Palmer, 1987).

The sharing of haplotypes between the two varieties indicates gene flow between the varieties and, as for the SSR marker cited above, gives support to Moreira *et al.* (2015) studies suggesting that cultivated and wild annatto are varieties of the same species. High pigment production in non–cultivated annatto was found where there are adjacent areas of intense cultivation of annatto, also suggesting gene flow from cultivated to wild by cross pollination and introgression. However, based on chloroplast variation it was not possible to trace the dispersal routes of annatto in the Brazilian territory.

REFERENCES

Akshatha, V., Giridhar, P. and Ravishankar, G.A. (2011) Morphological diversity in *Bixa orellana* L. and variations in annatto pigment yield. *J. Hortic. Sci. Biotechnol.* **86**, 319–324.

Almeida, F.O. de and Neves, E.G. (2015) Evidências arqueológicas para a origem dos tupi-guarani no leste da amazônia. *Mana* **21**, 499–525.

Andueza-Noh, R.H., Martínez-Castillo, J. and Chacón-Sánchez, M.I. (2015) Domestication of small-seeded lima bean (*Phaseolus lunatus* L.) landraces in Mesoamerica: evidence from microsatellite markers. *Genetica* **143**,657–669.

Bandelt, H.J., Forster, P. and Röhl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48.

Bitocchi, E., Bellucci, E., Giardini, A., *et al.* (2013) Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytol.* **197**, 300–313.

Brown, A.H.D. (2010) Variation under domestication in plants: 1859 and today. *Philos. Trans. R. Soc. B Biol.*

Sci. **365**, 2523–2530.

Cao, K., Zheng, Z., Wang, L., *et al.* (2014) Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biol.* **15**, 415.

Cavalli-Sforza, L.L. and Edwards, A.W.F. (1967) Phylogenetic Analysis Models and Estimation Procedures. *Am. J. Hum. Genet.* **19**, 233–57.

Clement, C.R. (1989) A Center of Crop Genetic Diversity in Western Amazonia. *Bioscience* **39**, 624–631.

Clement, C.R., Cristo-Araújo, M. de, D'Eeckenbrugge, G.C., Alves Pereira, A. and Picanço-Rodrigues, D. (2010) Origin and Domestication of Native Amazonian Crops. *Diversity* **2**, 72–106.

Clement, C.R., Rodrigues, D.P., Alves-Pereira, A., Mühlen, G.S., Cristo-Araújo, M. De, Moreira, P.A., Lins, J. and Reis, V.M. (2016) Crop domestication in the upper Madeira River basin. *Bol. do Mus. Para. Emilio GoeldiCiencias Humanas* **11**, 193–205.

Dieringer, D. and Schlötterer, C. (2003) microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* **3**, 167–169.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of Plant DNA from fresh tissue. *Focus (Madison).* **12**, 13–15.

Dubcovsky, J. and Dvorak, J. (2007) Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. *Science (80-.).* **316**, 1862–1866.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.

Erickson, C. (1995) Archaeological methods for the study of ancient landscapes of the Llanos de Mojos in the Bolivian Amazon. In Archaeology in the lowland American tropics: Current analytical methods and applications. (Stahl, P., ed), pp. 66–95. Cambridge: Cambridge University Press.

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**, 2611–20.

Excoffier, L. and Lischer, H.E.L. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567.

Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Seattle: University of Washington.

Fu, Y.. and Li, W.. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Gepts, P. (2004) Crop Domestication as a Long Term Selection Experiment, Oxford, UK: John Wiley & Sons, Inc.

Harlan, J. (1992) Crops and man, Madison: American Society of Agronomy.

Jakob, S.S. and Blattner, F.R. (2006) A Chloroplast Genealogy of Hordeum (Poaceae): Long-Term Persisting Haplotypes, Incomplete Lineage Sorting, Regional Extinction, and the Consequences for Phylogenetic Inference. *Mol. Biol. Evol.* **23**, 1602–1612.

Jombart, T. and Ahmed, I. (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R. and Dhawan, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177**, 309–334.

Kassa, M.T., Penmetsa, R.V., Carrasquilla-Garcia, N., Sarma, B.K., Datta, S., Upadhyaya, H.D., Varshney, R.K., Wettberg, E.J.B. von and Cook, D.R. (2012) Genetic Patterns of Domestication in Pigeonpea (*Cajanus cajan* (L.) Millsp.) and Wild Cajanus Relatives Shiu, S.-H., ed. *PLoS One* **7**, e39563.

Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version

7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874.

Ladizinsky, G. (1985) Founder effect in crop-plant evolution. *Econ. Bot.* **39**, 191–199.

Leal, F. and Clavijo, C.M. (2010) Acerca de la história, taxonomia, botánica y usos de *Bixa orellana* L. *Rev. Unell. Cienc. Tec* **1**, 78–86.

Levis, C., Costa, F.R.C., Bongers, F., *et al.* (2017) Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. *Science (80-.).* **355**, 925–931.

Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–2.

López-Gartner, G., Cortina, H., McCouch, S.R. and Moncada, M.D.P. (2009) Analysis of genetic structure in a sample of coffee (*Coffea arabica* L.) using fluorescent SSR markers. *Tree Genet. Genomes* **5**, 435–446.

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez G, J., Buckler, E. and Doebley, J. (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 6080–4.

Meyer, R.S. and Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852.

Moreira, P.A., Lins, J., Dequigiovanni, G., Veasey, E.A. and Clement, C.R. (2015) The Domestication of Annatto (*Bixa orellana*) from *Bixa urucurana* in Amazonia. *Econ. Bot.* **69**, 127–135.

Morrell, P.L. and Clegg, M.T. (2007) Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. *Proc. Natl. Acad. Sci.* **104**, 3289–3294.

Palmer, J.D. (1987) Chloroplast DNA Evolution and Biosystematic Uses of Chloroplast DNA Variation. *Source Am. Nat. Am. Nat* **13026**.

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.

Piperno, D.R. and Pearsall, D.M. (1998) The origins of agriculture in the lowland neotropics, Academic Press.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

R Core Team (2015) R: A language and environment for statistical computing.

Rong, J., Lammers, Y., Strasburg, J.L., Schidlo, N.S., Ariyurek, Y., Jong, T.J. de, Klinkhamer, P.G.L., Smulders, M.J.M. and Vrieling, K. (2014) New insights into domestication of carrot from root transcriptome analyses. *BMC Genomics* **15**, 895.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–25.

San-San-Yi, Jatoi, S.A., Fujimura, T., Yamanaka, S., Watanabe, J. and Watanabe, K.N. (2008) Potential loss of unique genetic diversity in tomato landraces by genetic colonization of modern cultivars at a non-center of origin. *Plant Breed.* **127**, 189–196.

Sandy-Cuen, P.M. and Becerra, R. (2003) Manejo campesino de recursos naturales. El achiote. *BioDiversitas* **7**, 7–11.

Santos, E.J.M. dos, Silva, A.L.S. da, Ewerton, P.D., Takeshita, L.Y. and Maia, M.H.T. (2015) Origins and demographic dynamics of Tupí expansion: a genetic tale. *Bol. do Mus. Para. Emílio Goeldi. Ciências*

Humanas **10**, 217–228.

Shaw, J., Lickey, E.B., Beck, J.T., *et al.* (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.* **92**, 142–66.

Shaw, J., Lickey, E.B., Schilling, E.E. and Small, R.L. (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *Am. J. Bot.* **94**, 275–88.

Staden, R., Beal, K.F. and Bonfield, J.K. (1998) The Staden package, 1998. *Methods Mol. Biol.* **132**, 115–130.

Tajima, F. (1989) Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Tanksley, S.D. (1997) Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. *Science (80-.).* **277**, 1063–1066.

Terrell, J.E., Hart, J.P., Barut, S., *et al.* (2003) Domesticated Landscapes: The Subsistence Ecology of Plant and Animal Domestication. *J. Archaeol. Method Theory* **10**, 323–368.

Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C. de F. (2016) Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **39**, 312–328.

Weir, B.S. (1996) Genetic data analysis II: methods for discrete population genetic data, Sinauer Associates.

Yang, S., Pang, W., Ash, G., Harper, J., Carling, J., Wenzl, P., Huttner, E., Zong, X. and Kilian, A. (2006) Low level of genetic diversity in cultivated Pigeonpea compared to its wild relatives is revealed by diversity arrays technology. *Theor. Appl. Genet.* **113**, 585–595.

Yano, O., Ikeda, H. and Hoshino, T. (2010) Phylogeography of the Japanese common sedge, *Carex conica* complex (Cyperaceae), based on chloroplast DNA sequence data and chromosomal variation. *Am. J. Bot.* **97**, 1365–76.

Zohary, D. (2004) Unconscious Selection and the Evolution of Domesticated Plants. *Econ. Bot.* **58**, 5–10.