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

Permalink
https://escholarship.org/uc/item/9q95912k

Journal
PloS one, 13(6)

ISSN
1932-6203

Authors
Dequigiovanni, Gabriel
Ramos, Santiago Linorio Ferreyra
Alves-Pereira, Alessandro
et al.

Publication Date
2018

DOI
10.1371/journal.pone.0198593

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

Gabriel Dequigiovanni¹, Santiago Linorio Ferreya Ramos², Alessandro Alves-Pereira¹, Eliane Gomes Fabri³, Doriane Picanço-Rodrigues⁴, Charles Roland Clement⁵, Paul Gepts⁶, Elizabeth Ann Veasey¹*

¹ Departamento de Genética, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo (ESALQ-USP), Piracicaba, São Paulo, Brasil, ² Instituto de Ciências Exatas e Tecnologia em Itacoatiara, Universidade Federal do Amazonas, Itacoatiara, Amazonas, Brasil, ³ Centro de Horticultura, Instituto Agronômico de Campinas, Campinas, São Paulo, Brasil, ⁴ Instituto de Ciências Biológicas, Universidade Federal do Amazonas (ICB-UFAM), Manaus, Amazonas, Brasil, ⁵ Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, Amazonas, Brasil, ⁶ Department of Plant Sciences, University of California, Davis, California, United States of America

* eaveasey@usp.br

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.
Introduction

Anatto (Bixa orellana L.) is a tropical American crop [1], which probably originated in Amazonia [2–4]. Anatto is commercially valuable due to its application in the food and cosmetics industries, as a natural dye to be used instead of synthetic ones [5]. 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.) [6,7], which belong 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 [4,8]. 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. excelsa and B. platycarpa are trees [7,9]. Ducke [10] hypothesized that B. excelsa might have been the wild ancestor of B. orellana, which was accepted by Schultes [4] and Meyer et al. [11]. However, B. excelsa is a tree and it is unlikely that domestication during the Holocene would transform all known populations into a shrub [9]. Analysis of the domestication syndrome in the shrubby Bixas allowed Moreira et al. [9] to propose that B. urucurana is the wild ancestor of cultivated annatto, B. orellana. They also accepted Pilger’s proposal, published by Kuntz [12], 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 [9].

B. orellana var. urucurana occurs in open forests and anthropogenic landscapes, although never cultivated and is always associated with riparian environments [7,9]. 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 [9]. The exact location where annatto was first domesticated is still unclear, mainly because of the wide distribution of variety urucurana in northern South America [9].

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 [13]. 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, generally neutral and highly polymorphic [14]. Hence, SSR markers are important tools to assess genetic diversity and genetic structure of populations, especially for wild species [14]. There are very few genetic studies in annatto [15–17] 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?, c) is genetic diversity geographically structured across these populations?, and d) what is the potential distribution of wild annatto in Amazonia? We used two approaches to answer these questions: neutral genetic variation (SSR markers) for questions ‘a’, ‘b’, ‘c’, and Ecological Niche Modeling (ENM) for question ‘d’ and how it may influence ‘a’, ‘b’ and ‘c’. ENM methods approximate a climatic 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 [18,19], and their possible influences on patterns of genetic diversity of populations [20,21]. 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 and also on online databases.

**Material and methods**

**Plant material**

During our field work from 2009 to 2015, 170 plants of wild annatto (*Bixa orellana* var. *urucurana*) were located and collected in 10 municipalities in the states of Rondônia, Pará and Roraima, in Brazilian Amazonia (Table 1, Fig 1). A variety of seed bearing fruits with different shapes were observed during the field collections, and plants were usually associated with riparian environments (S1 Fig). 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). No special permission was required for our sampling of annatto plants considering it was conducted according to the resolution 21, from CNPq (National Council for Scientific and Technological Development), in Brazil, which allows researchers to collect leaf samples for genetic analysis, as it is characterized as scientific research on phylogenetic relationships between geographic regions with annatto. Also, we have not accessed traditional knowledge related to wild annatto plants.

**DNA isolation, PCR amplification and genotyping of SSRs**

Total genomic DNA was extracted from young leaves following Doyle and Doyle [22] with CTAB 3%. DNA concentration was determined by comparison with known concentrations of standard DNA (lambda DNA, Invitrogen) during electrophoresis in agarose gels (1% (w/v)) stained with GelRed (Biotium) under ultraviolet light.

Sixteen SSR markers developed for *B. orellana* [17,23] were used in this study (Table 2; S1 Table). Three fluorescent dyes (NED, FAM and HEX) were attached to the 5’ end of the M13 universal primer sequence (*5’ - CACGACGTTGTAAAA CGAC-3’*) following Schuelke [24]. Polymerase chain reaction (PCR) amplification of the DNA samples was done in a MyCycler

<table>
<thead>
<tr>
<th>Population ID / Municipality, State*</th>
<th>N</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 –Cabaixi, RO</td>
<td>45</td>
<td>-13.4883</td>
<td>-60.60608</td>
</tr>
<tr>
<td>2 –Cerejeiras, RO</td>
<td>32</td>
<td>-13.1717</td>
<td>-60.80942</td>
</tr>
<tr>
<td>3 –Corumbiara, RO</td>
<td>26</td>
<td>-12.9915</td>
<td>-60.92277</td>
</tr>
<tr>
<td>4 –São Francisco do Guaporé, RO</td>
<td>9</td>
<td>-11.7261</td>
<td>-62.34804</td>
</tr>
<tr>
<td>6 –Ariqueles, RO</td>
<td>18</td>
<td>-9.92515</td>
<td>-63.07129</td>
</tr>
<tr>
<td>7 –Mucaiai, RR</td>
<td>5</td>
<td>2.370000</td>
<td>-61.44000</td>
</tr>
<tr>
<td>8 –Monte Alegre, PA</td>
<td>5</td>
<td>-1.98119</td>
<td>-54.16811</td>
</tr>
<tr>
<td>9 –Almeirim, PA</td>
<td>4</td>
<td>-1.24172</td>
<td>-53.04789</td>
</tr>
<tr>
<td>10 –Bom Jesus do Tocantins, PA</td>
<td>7</td>
<td>-5.10388</td>
<td>-48.54889</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>170</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* PA–Pará, RO–Rondônia, RR–Roraima

https://doi.org/10.1371/journal.pone.0198593.t001
Thermal Cycler (Bio-Rad), and performed in a final volume of 10 μL, consisting of 20 ng of DNA template, 1X PCR buffer (Fermentas, Vilnius, Lithuania), 0.25 mM of each dNTP, 1.5 mM of MgCl₂, 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 [24] 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 (Table 2) 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% (w/v)) 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).

SSR data analysis
Possible clusters of wild annatto were simulated using a Bayesian analysis with Structure software [25]. 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 *adhoc* method of Evanno et al. [26].

Based on the original populations and the clusters identified by *Structure*, we estimated parameters for genetic diversity, including number of alleles (*A*), effective allele number (*Ne*), allelic richness (*A*R) [27], observed (*H*O) and expected heterozygosities (*H*E), in addition to Wright’s [28] inbreeding coefficient (*f*). The apparent outcrossing rate (*^t_a*^) was estimated considering the inbreeding coefficient (*f*) for each population [29], so that *^t_a*^ = (1—*f*)/ (1+*f*). The genetic diversity indices *A*, *Ne*, *H*O and *H*E were estimated with GenAlEx 6.5 [30], and estimations of *A*R and *f*, with confidence intervals based upon 1,000 bootstrap replicates, were obtained with diversity [31] and poppr [32] for R [33].

In order to represent the relationships between individuals and populations, neighbor-joining [34] dendrograms were constructed with Phylip 3.5 [35], based on Cavalli-Sforza and Edwards’chord distance (*D*CE) [36] obtained with MSA 4.05 [37]. The chord distance is a geometric distance and performs well for the reconstruction of relationships among populations [38]. 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 [30].

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 [30]. Gene flow (*N*m) among populations was estimated by calculating *N*m = (1—*F*ST)/4*F*ST [39]. In addition, the Mantel test was used to evaluate the correlation between Nei’s genetic distance and geographic distance (km) among populations using Adegenet [40] for R [33]. Significance was assessed by conducting 9,999 permutations.

**Table 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) heterozgyosities, and Shannon’s information index (*I*).

<table>
<thead>
<tr>
<th>Loci</th>
<th>A</th>
<th><em>H</em>O</th>
<th><em>H</em>E</th>
<th><em>I</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>BorA2</td>
<td>5</td>
<td>0.354</td>
<td>0.438</td>
<td>0.712</td>
</tr>
<tr>
<td>BorA3</td>
<td>17</td>
<td>0.391</td>
<td>0.463</td>
<td>0.916</td>
</tr>
<tr>
<td>BorA5</td>
<td>12</td>
<td>0.401</td>
<td>0.525</td>
<td>0.968</td>
</tr>
<tr>
<td>BorB1</td>
<td>17</td>
<td>0.311</td>
<td>0.641</td>
<td>1.268</td>
</tr>
<tr>
<td>BorB4</td>
<td>17</td>
<td>0.352</td>
<td>0.609</td>
<td>1.111</td>
</tr>
<tr>
<td>BorB5</td>
<td>12</td>
<td>0.379</td>
<td>0.458</td>
<td>0.834</td>
</tr>
<tr>
<td>BorB12</td>
<td>14</td>
<td>0.421</td>
<td>0.508</td>
<td>1.015</td>
</tr>
<tr>
<td>BorC5</td>
<td>12</td>
<td>0.668</td>
<td>0.712</td>
<td>1.286</td>
</tr>
<tr>
<td>BorD1</td>
<td>8</td>
<td>0.357</td>
<td>0.423</td>
<td>0.752</td>
</tr>
<tr>
<td>BorD2</td>
<td>10</td>
<td>0.468</td>
<td>0.434</td>
<td>0.841</td>
</tr>
<tr>
<td>BorF9</td>
<td>12</td>
<td>0.231</td>
<td>0.629</td>
<td>1.220</td>
</tr>
<tr>
<td>BorG4</td>
<td>14</td>
<td>0.465</td>
<td>0.569</td>
<td>1.102</td>
</tr>
<tr>
<td>BorG11</td>
<td>19</td>
<td>0.426</td>
<td>0.679</td>
<td>1.354</td>
</tr>
<tr>
<td>BorH3</td>
<td>4</td>
<td>0.158</td>
<td>0.292</td>
<td>0.472</td>
</tr>
<tr>
<td>BorH7</td>
<td>5</td>
<td>0.494</td>
<td>0.443</td>
<td>0.692</td>
</tr>
<tr>
<td>BorH10</td>
<td>16</td>
<td>0.239</td>
<td>0.540</td>
<td>1.066</td>
</tr>
<tr>
<td>Total</td>
<td>194</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mean</td>
<td>12.125</td>
<td>0.382</td>
<td>0.522</td>
<td>0.975</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0198593.t002
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 [41]. Maxent estimates the potential distribution of a taxa from a maximum entropy probability distribution using presence-only data [42]. 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’ [43]. 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.

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 [44]. 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 [45] for R [33].

**Results**

**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 2). The observed heterozygosity ($H_O$) 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 2).

Genetic diversity estimates of the 10 populations showed a mean number of alleles per locus ($\bar{A}$) of 3.86 (Table 3). Forty-one private alleles were observed, representing 21% of all alleles. The population from Bom Jesus do Tocantins (Population 10 in Table 1 and Fig 1) showed the highest number of private alleles (15) (Table 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 \((t_a)\) was 0.609. When disregarding the populations with small sampling sizes, such as Monte Alegre \((N = 5)\), Mucajaí \((N = 5)\) and Almeirim \((N = 4)\), the mean value of this parameter increased to 0.690.

### Genetic structure

The 170 wild annatto plants of 10 populations were grouped into genetic clusters by the Structure simulations, with a clear \(\Delta K\) maximum at \(K = 3\), and possible substructure at \(K = 2\) and \(K = 7\) (Fig 2, S2 Fig). 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 Ji-Paraná, the Ji-Paraná 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 (Fig 1). At \(K = 2\), South RO was allocated in one group and Central RO was clustered together with populations from PA and RR. At \(K = 7\), the PA and RR

### Table 3. Genetic diversity estimated for 10 populations of *Bixa orellana* var. urucurana, including mean number of alleles per locus \((A)\), allelic richness \((A_R)\), mean number of effective alleles per locus \((N_E)\), observed \((H_O)\) and expected \((H_E)\) heterozygosities, inbreeding coefficient \((f = 1 – \frac{H_O}{H_E})\), and apparent outcrossing rate \((t_a)\).

<table>
<thead>
<tr>
<th>Population, State</th>
<th>(A) (private alleles)</th>
<th>(A_R)</th>
<th>(N_E)</th>
<th>(H_O)</th>
<th>(H_E)</th>
<th>(f)</th>
<th>(t_a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cabixi, RO</td>
<td>5.813 (1)</td>
<td>1.582</td>
<td>3.000</td>
<td>0.449</td>
<td>0.577</td>
<td>0.217</td>
<td>0.643</td>
</tr>
<tr>
<td>2. Cerejeiras, RO</td>
<td>4.875 (1)</td>
<td>1.558</td>
<td>2.845</td>
<td>0.442</td>
<td>0.554</td>
<td>0.181</td>
<td>0.693</td>
</tr>
<tr>
<td>3. Corumbiara, RO</td>
<td>4.750 (1)</td>
<td>1.577</td>
<td>3.127</td>
<td>0.445</td>
<td>0.571</td>
<td>0.196</td>
<td>0.672</td>
</tr>
<tr>
<td>4. S. F. do Guaporé, RO</td>
<td>3.625 (3)</td>
<td>1.564</td>
<td>2.512</td>
<td>0.437</td>
<td>0.548</td>
<td>0.168</td>
<td>0.712</td>
</tr>
<tr>
<td>5. Ji-Paraná, RO</td>
<td>5.000 (3)</td>
<td>1.645</td>
<td>3.379</td>
<td>0.426</td>
<td>0.636</td>
<td>0.309</td>
<td>0.528</td>
</tr>
<tr>
<td>6. Ariquemes, RO</td>
<td>3.875 (3)</td>
<td>1.565</td>
<td>2.566</td>
<td>0.453</td>
<td>0.553</td>
<td>0.166</td>
<td>0.715</td>
</tr>
<tr>
<td>7. Mucajaí, RR</td>
<td>3.438 (9)</td>
<td>1.631</td>
<td>2.816</td>
<td>0.268</td>
<td>0.596</td>
<td>0.565</td>
<td>0.278</td>
</tr>
<tr>
<td>8. Monte Alegre, PA</td>
<td>1.813 (2)</td>
<td>1.338</td>
<td>1.653</td>
<td>0.313</td>
<td>0.321</td>
<td>0.047</td>
<td>0.910</td>
</tr>
<tr>
<td>9. Almeirim, PA</td>
<td>2.063 (3)</td>
<td>1.353</td>
<td>1.865</td>
<td>0.146</td>
<td>0.327</td>
<td>0.508</td>
<td>0.326</td>
</tr>
<tr>
<td>10. BJTocantins, PA</td>
<td>3.313 (15)</td>
<td>1.535</td>
<td>2.596</td>
<td>0.471</td>
<td>0.516</td>
<td>0.081</td>
<td>0.850</td>
</tr>
<tr>
<td>Mean</td>
<td>3.856</td>
<td>1.535</td>
<td>2.636</td>
<td>0.385</td>
<td>0.519</td>
<td>0.243</td>
<td>0.609</td>
</tr>
</tbody>
</table>

*PA–Pará, RO–Rondônia, RR—Roraima
* Significant based upon 1,000 bootstrap replicates

https://doi.org/10.1371/journal.pone.0198593.t003

---

Fig 2. Assignment of each of 170 *Bixa 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.

https://doi.org/10.1371/journal.pone.0198593.g002
A 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 within both groups (Table 3). Also, a separate group was formed at Corumbiara, in southern Rondônia, where we found mixed populations from South and Central Rondônia at both $K = 2$ and $K = 3$ (Fig 2).

Relationships among populations in the dendrogram (Fig 3) generally agreed with Structure and PCoA (not shown data). The relationship among individuals in the dendrogram (S3 Fig) also agreed with Structure and PCoA results. In the dendrogram of individual plants (S3 Fig), 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.

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.

Fig 3. Unrooted neighbor-joining dendrogram of 10 *Bixa orellana* var. *urucurana* populations collected in Brazilian Amazonia based on Cavalli-Sforza & Edwards chord distance [36] estimated from 16 SSR. Colors are according to the probability of occurrence of *Bixa orellana* var. *urucurana* simulated by Ecological Niche Modeling (Fig 5). Red = very high probability, yellow = high probability and green = moderate probability.

https://doi.org/10.1371/journal.pone.0198593.g003
AMOVA revealed that 21% of the genetic variation was among populations, and the remaining 79% was within populations (p = 0.000), showing that, although most of the diversity is within populations, the variation due to sub-division of the populations is significant (Table 4). 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 ($N_m = 4.843$).

**Species distribution modelling and climatic adaptation**

The result of the PCA 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 differentiation between populations from Rondônia, and Pará and Roraima along the second PCA axis (Fig 4). 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 (S2 Table 4. Analysis of molecular variance (AMOVA) performed for 16 SSR and 170 *Bixa orellana var. urucurana* plants collected in Brazilian Amazonia.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>Variance</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among populations</td>
<td>9</td>
<td>415.758</td>
<td>46.195</td>
<td>1.301</td>
<td>21</td>
</tr>
<tr>
<td>Within populations</td>
<td>330</td>
<td>1611.165</td>
<td>4.882</td>
<td>4.882</td>
<td>79</td>
</tr>
<tr>
<td>Total</td>
<td>339</td>
<td>2026.924</td>
<td>6.183</td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0198593.t004

![Fig 4. 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 at K = 3 and online databases: Pink = South Rondônia accessions; Blue = Central Rondônia accessions; Green = accessions from the states of Pará and Roraima; and Gray = online databases occurrences.](https://doi.org/10.1371/journal.pone.0198593.g004)
Principal component 1 (PC1) represented 71% of the variation and was mostly explained by variable Bio12 (Annual Precipitation). The annual precipitation varied from 1,309 mm (on the left) to 3,644 mm (on the right). Our sampled populations presented annual precipitations ranging from 1,669 mm to 2,192 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 (S2 Table), ranging from 87 mm (on the top) to 1,388 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.

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 (Fig 5). The climate envelope of wild annatto is largely determined by precipitation, and the most important 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.

**Discussion**

**Genetic diversity**

This is the first genetic study with wild populations of annatto. The levels of heterozygosity averaged over all loci (*H*<sub>E</sub> = 0.520; *H*<sub>O</sub> = 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 [46]. 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 [47] and generally further bottlenecks due to distribution and diversification [48], including different selection pressures for yield [49].

The high levels of inbreeding coefficient (\(f\)) 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 wild annatto. The existence of crosses between related individuals increases homozygosity [50]. These results are in agreement with other studies on cultivated annatto, such as those obtained by Rivera-Madrid et al. [51], 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. [52] also concluded that there is natural selfing in annatto. Valdez-Ojeda et al. [16] found high multilocus outcrossing rates (0.75) based on 50 SRAP markers and concluded that \(B. \) orellana has a mixed mating system. A recent study compared the mating system of annatto under different agronomic systems in Mexico using SRAP markers [51], showing a mixed mating system for annatto cultivated in backyards, while those cultivated under polyculture (milpa system) and monoculture systems showed predominantly outcrossing mating systems. Considering that most of the above results have shown mixed mating system for cultivated annatto, we may conclude that domestication of annatto did not include modifications in the mating system.

**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 [13]. 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. [53] 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 [53].

The mean level of gene flow among \(B. \) orellana var. urucurana populations was low in this study (\(N_m = 0.545\)), since many of the populations were very distant apart, especially those from Rondônia and the ones from Roraima and Pará, but also between those from Roraima and Pará. According to Wright [54], 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 dispersal along rivers and streams [9], but also through cross-pollination by bees [55]. 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 (see discussion...
below). 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 Ji Parana) are in the Ji-Paraná River basin, so that gene flow by the transport of seeds along the rivers between these populations is infrequent. However, mixed populations between the two groups were found at Corumbiara, in southern Rondônia, classified as a separate group at $K = 7$, indicating an intermediate area with a higher gene flow between the two groups.

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 [36] chord distance and the neighbor-joining method, as well as in the PCoA and the Bayesian analysis, three genetically distinct and consistent groups were identified. The groups were formed based on the geographical location of the collected populations. Carvalho et al. [56], using isozymes, also found that genetic differences in cultivated annatto accessions correspond to distinct geographical locations. However, Medina et al. [57] 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. [58] 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. These 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 may be 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. [9] 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 Amazonia, and in most cases, in anthropogenic areas [9], 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.

**Distribution and climatic adaptation of wild annatto populations**

The high global $F_{ST}$ value (0.201), identifying strong structure among populations, is partially due to our sampling effort, but may also reflect the real and modeled distribution of wild annatto. According to the ENM, wild annatto is not uniformly distributed throughout Amazonia, with a large area in Central Amazonia (Fig 5) unsuitable for the species. If this ENM is accurate, it may be very difficult to find large natural populations of wild annatto in this wide area. This fragmented distribution may be a major barrier for gene flow among populations, which may explain the high genetic differentiation and also the pattern of isolation by distance suggested by the Mantel test. This model also helps explain the lack of observation of wild annatto in these areas during our field collections, or the observation of smaller populations, and the low number of herbaria registers. Comparing the Southern Rondonia populations with the Central Rondonia populations, there were an abundance of plants to be sampled in Southern Rondônia, while the wild annatto populations found in Central Rondônia had fewer plants. As a matter of fact, we collected all of the plants from the center of Rondônia.
populations that we came across with. South Rondônia is an area of high suitability for wild annatto (Fig 5), according to the ENM, which may explain the higher sampling obtained in this area.

In addition, we may consider that the high divergence among populations may be due to what is known as the “arc of deforestation”, a region where the agricultural frontier advances towards the forest [59]. There are 500 thousand km² of land that goes from the east and south of Pará towards the west, passing through the States of Mato Grosso, Rondônia and Acre (Fig 1). The arc of deforestation has the highest rates of deforestation of the Amazon forest, which is also probably causing the fragmentation of wild annatto populations. The possible occurrence of small isolated populations of wild annatto across Amazonia also agrees with the higher levels of intrapopulation inbreeding coefficients, suggesting the action of genetic drift coupled with inbreeding within the populations included in this study.

Temperature and precipitation are considered as major factors in determining species distributions [60]. Our ENM model suggests 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 suggest the species tolerance to 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 identified as major selective pressures driving plant adaptation because they are very important for plant growth, development, and reproduction [61,62]. 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 [63].

Not surprisingly, the potential distribution coincided approximately with the current occurrence reported in online distribution databases, but also suitable localities were predicted outside the presently known range of the species. These localities could be targeted with field surveys that might identify unknown populations. However, a considerable number of occurrences 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. [3], 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 [64]. Levis et al. [65] also found higher abundance and richness of domesticated species in southwestern Amazonia.

Piperno and Pearsall [66] 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 [67]. 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 [67]. Complex societies inhabited this region at the time of the European conquest, and managed dozens of species, leading Clement [68] 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 2,400 years before present [69].
The Madre de Dios Basin is also a highly suitable area for *B. orellana* var. *urucurana*. According to Leal and Clavijo [70], 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 [64]. The Madre de Dios Basin drains an area of approximately 90,000 km along the eastern flank of the Andes in southeastern Peru, ranging in elevation from 200 m to over 4,000 m [71]. The vegetation is predominantly evergreen or semi-evergreen forest [72]. It presents a humid tropical climate with annual rainfall varying from 1,200 mm to 3,300 mm, generally increasing from east to west, and the rainy season occurring from October to April [72].

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 annual temperatures ranging from 26˚C to 28˚C and monthly average maximum temperatures between 34˚C and 37˚C. The rainfall of the Llanos region shows a regime characterized by very pronounced differences during the months of the year, with annual rainfall ranging from 850 mm to 1,800 mm [73].

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 drier or seasonally drier areas, and are located in the peripheries of the Amazon, consistent with the favorable areas identified by Moreira et al. [9]. 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 Ecuador 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. *urucurana* to different niches.

According to our ENM model, we noticed that the distribution of our wild annatto sampling was made in quite marginal populations (suitability for Central Rondônia, Pará and Roraima is less than 0.7) and therefore, the picture of the genetic diversity is quite partial. We also observed that in the area between the two regions, there are no predicted populations (low suitability). Therefore, the gene flow among these two different regions could not be recent, but historical, and also probably due to some other historical factors of the populations. We do not know if there are populations in Ecuador that could clarify the relationship among the two main regions studied. The ENM model results in this study implies the recommendation that further collection expeditions should be made sampling *B. orellana* var. *urucurana* populations from eastern Ecuador, western Venezuela, northeastern Colombia and the Llanos de Mojos, Bolivia, as well as the State of Mato Grosso and Northeast Brazil.

**Conclusions**

The microsatellite loci used in this study revealed high levels of genetic diversity in 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. Interestingly, our ENM predicted a wide area of low suitability for wild annatto across Central Amazonia. This predicted occurrence plus increasing population fragmentation resulting from Amazonia deforestation contribute to the low genetic connectivity among disjunct populations of wild annatto. Therefore, our study demonstrates how ecological and anthropic factors may have an impact on the genetics of a native Amazonian species. 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.

Supporting information

S1 Table. SSR data for each *Bixa orellana* var. *urucurana* individual and populations. (TXT)

S2 Table. 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). (DOCX)

S1 Fig. Morphological variation on fruits and plant architecture of wild annatto (*Bixa orellana* var. *urucurana*). (TIFF)

S2 Fig. 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. (TIFF)

S3 Fig. 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. (TIFF)

Acknowledgments

This project was financially supported by the São Paulo Research Foundation (FAPESP) (2011/10357-8; 2012/08307-5) and by the National Council for Scientific and Technological Development (CNPq) CT Amazônia (575588/2008-0). G. D. was supported by FAPESP with PhD scholarships (2013/08884-5; 2016/05912-6). A.A.P. thanks the Coordination for the Improvement of Higher Education Personnel (CAPES) for a post-doctoral scholarship. E.A.V. and C.R.C. were supported by CNPq research fellowships.

Author Contributions

**Conceptualization:** Gabriel Dequigiovanni, Charles Roland Clement, Paul Gepts, Elizabeth Ann Veasey.

**Data curation:** Gabriel Dequigiovanni, Santiago Linorio Ferreyra Ramos.

**Formal analysis:** Gabriel Dequigiovanni, Santiago Linorio Ferreyra Ramos, Alessandro Alves-Pereira.

**Funding acquisition:** Charles Roland Clement, Elizabeth Ann Veasey.


Project administration: Gabriel Dequigiovanni, Elizabeth Ann Veasey.

Resources: Gabriel Dequigiovanni, Charles Roland Clement.

Software: Gabriel Dequigiovanni, Santiago Linorio Ferreyra Ramos, Alessandro Alves-Pereira.


Visualization: Gabriel Dequigiovanni.


Writing – review & editing: Gabriel Dequigiovanni, Alessandro Alves-Pereira, Charles Roland Clement, Paul Gepts, Elizabeth Ann Veasey.

References


18. Peterson AT. Predicting the geography of species’ invasions via ecological niche modeling. Q Rev Biol. 2003; 78: 419–33. PMID: 14737826


58. de Menezes IPP, Gaiotto FA, Hoffmann LV, Ciampi AY, Barros o PAV. Genetic diversity and structure of natural populations of *Gossypium mustellinum*, a wild relative of cotton, in the basin of the De Contas River in Bahia, Brazil. *Genetica*. 2014; 142: 99–108. [https://doi.org/10.1007/s10709-014-9757-6](https://doi.org/10.1007/s10709-014-9757-6) PMID: 24473734


