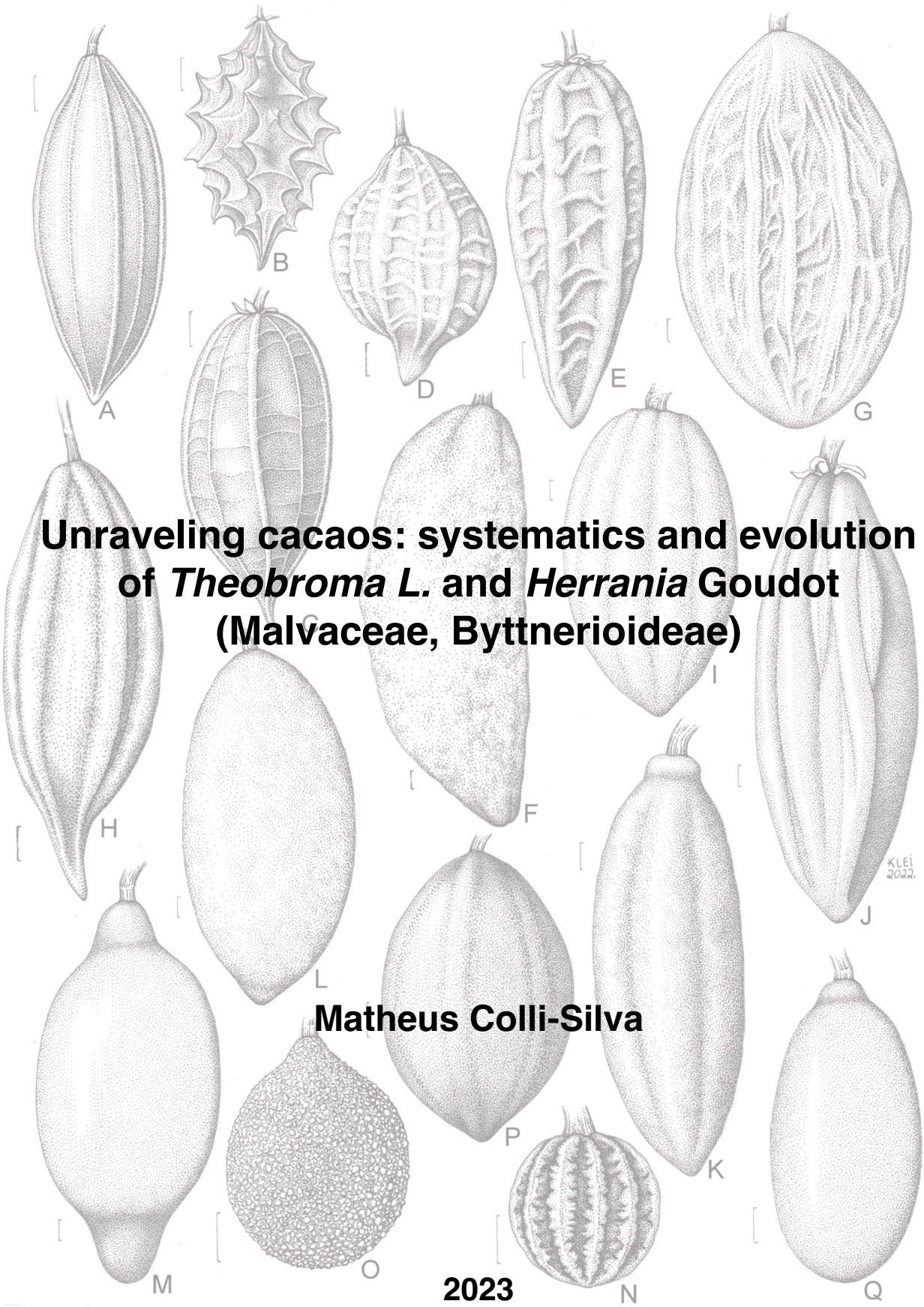

Unraveling cacaos: systematics and evolution of *Theobroma L.* and *Herrania* Goudot (Malvaceae, Byttnerioideae)

Matheus Colli-Silva

2023

UNIVERSITY OF SÃO PAULO
INSTITUTE OF BIOSCIENCES

MATHEUS COLLI SILVA

Unraveling cacaos: systematics and evolution of *Theobroma* L. and *Herrania* Goudot (Malvaceae, Byttnerioideae)

Dissertation submitted to the Institute of Biosciences of the University of São Paulo for the degree of **Doctor in Science**, area of **Botany**.

Advisor: Prof. Dr. José Rubens Pirani
Co-advisor: Prof. Dr. James E. Richardson

São Paulo

2023

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE BIOCIÊNCIAS

MATHEUS COLLI SILVA

Desbravando os cacaus: sistemática e evolução de *Theobroma* L. e *Herrania* Goudot (Malvaceae, Byttnerioideae)

Tese apresentada ao Instituto de Biociências da Universidade de São Paulo para a obtenção de Título de **Doutor em Ciências** na área de **Botânica**.

Orientador: Prof. Dr. José Rubens Pirani
Coorientador: Prof. Dr. James E. Richardson

São Paulo

2023

Colli-Silva, Matheus.
Desbravando os cacaus: sistemática e evolução de *Theobroma* L. e *Herrania* Goudot (Malvaceae, Byttnerioideae) / Matheus Colli Silva; Orientador José Rubens Pirani.—São Paulo, 2023.
404 p.

Tese (Doutorado—Programa de Pós-Graduação em Botânica) - Instituto de Biociências, Universidade de São Paulo, 2023.

1. Amazônia 2. Fanerógamas 3. Filogenia. 4. Malvales. I. Universidade de São Paulo. Instituto de Biociências. Departamento de Botânica.

Comissão julgadora:

Dr(a).

Dr(a).

Dr(a).

Dr. **José Rubens Pirani** (Orientador)

ACKNOWLEDGMENTS

This work was supported by Brazil's Coordination for the Improvement of Higher Education Personnel (CAPES)—Financial Code 001. I am also grateful to FAPESP—The São Paulo Research Foundation—for the massive support approved towards the conduction of my research (Grants 2020/01375-1, 2020/10206-9 and 2021/08635-1).

I started my PhD in an absolutely atypical and unfortunate period: the pandemic. Sometimes I cannot believe that even under such terrible circumstances I could produce a thesis which, honestly, I am absolutely proud of. This would not have been possible without the supervision of my dear mentor Prof. Pirani, who has followed my whole academic journey even earlier than when I joined USP's Lab of Systematics, Evolution and Biogeography of Vascular Plants, when I still attended to undergraduate disciplines. After almost ten years working together, from my Scientific Initiation, then during my masters' and now in the PhD, I will be forever grateful for his supervision, support, friendship and boundless belief in my work. We made a lot together so far towards the study of cacaos, the Malvaceae, biogeography and the Sapindales, me as a student and him as a supervisor. Now, I am absolutely confident that we will keep being intensively collaborating in the near future, now in the condition of collaborators.

I have a number of professors and prominent researchers who also helped me during my PhD in many ways, either by criticizing and giving ideas on many aspects of the project, or by enabling the realization my research. I am absolutely grateful to Profs. Lucia Lohmann, Renato Mello-Silva (*in memoriam*), Paulo Sano, Rafaela Forzza, Jefferson Prado, Juliana El-Ottra, Gregorio Ceccantini, Inês Cordeiro, Marcelo Pace, Alison Nazareno, Alex Antonelli, Marcia Rizzutto, Mike Hopkins, André Gil and Pedro Viana, Julio Betancur, Eduardo Gomes Neves, Jennifer Watling, and many others that made myself the scientist I am today.

Many thanks to Prof. James E. Richardson, who also agreed to supervise me in the condition of co-advisor, since 2021. James has always been excited about my research and very proactive in helping to establish a global network of scientists who study cacao topics. Ever since my first contact with him, he already inserted me in the field of cacao studies and crop genomics, which included me to know my dear colleague Dr. Ana Maria Bossa-Castro, from Los Andes University, who worked with me intensively, especially on the phylogeny of *Theobroma*. Thanks to both of you for the support and leadership!

I think three other names deserve a special mention: Fabián Michelangeli, Antonio Figueira and Laurence Dorr. Fabián agreed to supervise me during my stay at the New York Botanical Garden as a short-term scholar; even though he did not work with cacaos, he received me with very enthusiasm and thrill to assist in some aspects of the taxonomic part of the project, which included the description of three new species of *Theobroma*. Fabián also enabled new contacts with many American institutions and personnel, not to mention that he made my stay in NYC much more pleasant.

Prof. Antonio, like James, ever since from the beginning of this project, has been always followed me and helped me, particularly with respect on the taxonomy, morphology and genetics of wild *T. cacao* populations. Prof. Antonio opened my mind in many aspects on how to integrate the agronomic field with botany. Thanks!

Finally, Larry Dorr, one of the greatest malvologists of our generation. I had the opportunity to meet Larry in person in late 2022, in NY, because of Fabián. In there, Larry already showed his notes on *Theobroma* he had been accumulating for decades; he did not hesitate in collaborate with me towards tricky cases of nomenclature of *Cacao sylvestris* and *C. guianensis*, and also helped me in some aspects of the taxonomic revision as well. Thanks, Larry, for trusting in my work and for mentoring me so much about the *Theobromas*!

I am absolutely honored and happy to know my "little littles" from NYC: Profa. Almecina Ferreira ("Mel"), Laura Affonso (Laurita), Beatriz Valente, Silvana Monteiro ("Sil"). What a great time together, little ones! I miss you all! Also, many thanks to Doug Daly, Jackie Kallunki, Wayt Thomas, Benjamin Torke, Sérgio ("Sergito") Guzman, Andrés Ávila and others that made my stay in NYC so pleasant and fun, even during the Christmas and New Years' Eve, where we all spent together!

My best regards and thanks for my dearest friends from the "Malvaceae" team of Brazil: Carlos D.M. Miranda, Maria Tereza Costa (Tê), Vânia Nobuko, Thales Coutinho, Marilia Duarte, Massimo Bovini, Aluisio Fernandes-Junior and others. What a great team! Especially Carlos, Tê and Vânia, who went to the field with me and explored the deep jungles of Amazonia for over one month: thanks for being with me in the best field expedition I made in my life so far!

I am grateful also to all of my colleagues and friends from the Systematics Lab. It's been a pleasure to be with you all during my stay at USP, all the "bandeijões" we went together, all the "fofocas" we shared, all the "cafés" we took in the morning and after lunch. Thanks Andressa Cabral, Luana Sauthier, Gisele Alves, Guilherme "Piranha" Antar, Marcelo Kubo, Augusto Giaretta, Rebeca Gama, Jennifer Lopes, Sandra Reinales (Sandrita, my lab partner!), Renato Ramos, Marcelo Devecchi, Jéssica N. Francisco, Luiz Fonseca, Maila Beyer, Annelise Frazão, Edu Lozano, Roberto "Mão" Baptista, Renato (Renatinho) Magri, Lucas Borges de Lima, Marco Pellegrini & Rafael de Almeida, Mirian "Mirtilo" Antonicelli, Italo "Gandhi" Freitas, Raquel "Phoebs" Bastos, Guilherme Paschoalini, Danilo Zavatin, Elton John de Lírio. And the list goes on and on... Love you all!

I also cannot forget about every single person from USP's staff team (Conceição, Zé Vitório, Vivi, Robertinha, Adriana, Eli, Fabrício) for the support during my stay at the University, especially after 2022, where we came back *in situ*.

I also would like to stress my gratitude to Renato Ramos, Paulo Sano and the Plano de Ação Territorial Espinhaço Mineiro with sources from the Brazilian Environment Ministry (MMA – ProEspécies Initiative, www.proespecies.eco.br) for making a cluster available for use, which was essential for me to develop most of my genomic analyses.

Last but not least, I am grateful to all my beloved family and friends for the support and for believing through all my choices. You will always have a place in my heart.

Thankful list is endless. This PhD was many things, but one thing for sure is that it was NOT made only by me, but by the help of many of the people mentioned above. My many thanks to you all. I am absolutely pleased and happy to cross every single one of you. I'm sure this is just the beginning of a wonderful journey on cacao and relatives, crop genomics, Malvaceae and etc.!

RESUMO

A Amazônia é a maior e mais rica floresta tropical do mundo, fonte de uma grande quantidade de espécies de plantas e o principal repositório de serviços ecossistêmicos relevantes globalmente. A floresta é também o berço de muitas espécies vegetais nativas da região Neotropical, e uma quantidade significativa delas desempenhou um importante papel na história da humanidade desde o início do Holoceno. Estudar a origem e a história natural desses grupos é, portanto, a base para construir uma compreensão holística sobre a Amazônia, e o "grupo dos cacaus", com espécies tradicionalmente alocadas em dois gêneros – *Theobroma* e *Herrania* – mostra-se como um ótimo modelo de estudo. Neste projeto de doutorado, propus revisar o estado nomenclatural e taxonômico das espécies silvestres de cacaus, estudando a evolução morfológica e genômica populacional de táxons selecionados. Especificamente, os objetivos desta tese são (1) revisar morfologia e distribuição dos táxons de *Theobroma* e *Herrania* atualmente reconhecidos, reavaliando as delimitações taxonômicas a nível específico; (2) buscar sinapomorfias morfológicas e discutir a evolução de tais caracteres; (3) avaliar como a diversidade genômica de populações de cacau (*T. cacao*) se comporta em relação aos seus níveis de divergência de genes sob seleção; e (4) explorar a origem e história natural do *cupuaçu* (*T. grandiflorum*), uma espécie economicamente relevante no Brasil, conhecida por seus frutos enormes e apreciada por sua polpa. Esta tese está organizada em nove capítulos e quatro partes: "tratamento taxonômico", "filogenética e evolução", "genômica populacional" e "mobilização de dados e estudos derivados", cada parte abordando um ou mais objetivos associados a este projeto. Primeiramente, meus resultados para a revisão taxonômica consistem em descrições completas e atualizadas para um total de 35 espécies de *Theobroma/Herrania* alocadas em seis seções, além de três espécies novas. Em segundo lugar, uma nova filogenia revelou a parafilia de *Theobroma* e subsidiou o restabelecimento de *Theobroma* incluindo *Herrania* como uma de suas seções. Em terceiro lugar, usando uma abordagem de sequenciamento de alto rendimento, pude rastrear os efeitos diferenciais de seleção natural/artificial em diferentes populações de cacau. Além disso, demonstrei pela primeira vez que o *cupuaçu* não uma espécie selvagem, mas sim uma forma domesticada criada a partir de seu parente próximo, o cupuí (*T. subincanum*), de modo que eu pude não apenas rastrear onde e quando o *cupuaçu* se originou, mas também mostrar a existência de um único evento de domesticação ocorrido no Médio-Alto Amazonas no Holoceno Médio, mediado por povos indígenas que manipulavam a floresta muito antes da colonização europeia. Finalmente, apresento um banco de dados de coleções de espécimes preservados de *Theobroma*, e descrevo um estudo que discute a área nativa de *T. cacao*, sob uma perspectiva de sensoriamento remoto. Os resultados desta tese não apenas impactam profundamente o entendimento sobre as espécies de cacaus nativas, a evolução da flora e dos povos da Amazônia, mas iluminam uma série de aspectos acerca da origem e história natural dos parentes silvestres de espécies cultivadas.

Palavras-chave: Amazônia, domesticação, seleção natural, filogenética, genômica populacional, taxonomia.

ABSTRACT

Amazonia is the largest and most biodiverse rainforest of the world, source of a large amount of plant species, and the main repository of ecosystem services relevant globally. The forest is the cradle of many plants from Tropical Americas, and a significant amount of these have played a major role in human history since early Holocene. Studying the origin and history of such groups is, therefore, the basis to build an integrative understanding of Amazonia, and the "cacao group", with species traditionally allocated to two genera—*Theobroma* and *Herrania*—is one such example. In this PhD project, I revisited the nomenclatural and taxonomic status of the wild cacao species, associated to an assessment of morphological, evolutionary and populational genomics of selected taxa. Specifically, I aimed at (1) revisit the morphology and distribution of current taxa, reevaluating taxonomic delimitations at species levels; (2) look for morphological synapomorphies and discuss how the evolution of such characters might have occurred; (3) evaluate how genetic variation of populations of cacao (*T. cacao*) behave regarding its divergence levels of genes under selection; and (4) depict the origin and geographic history of *cupuaçu* (*T. grandiflorum*), a tree crop economically relevant in Brazil, known for its humongous fruit and appreciated for its pulp. This dissertation is organized in nine chapters and four parts: "taxonomic treatment", "phylogenetics and evolution", "populational genomics" and "data mobilization and by-product studies", each one tackling one or more goals associated to this project. First, results for the taxonomic revision consist of complete and updated descriptions of a total of 35 species of *Theobroma/Herrania* divided into six sections, including three new species here described. Second, a new phylogeny revealed the paraphyly of *Theobroma*, all of which subsidized a reestablishment of *Theobroma* including *Herrania* as a section of the first. Third, using a high-throughput sequencing approach, I could track differential effects of selection on different populations of cacao. Additionally, for the first time I demonstrated that *cupuaçu* is actually not a natural species, but a domesticated form selected from a wild close relative, *cupuí* (*T. subincanum*), so that I could not only track where and when this happened, but also pinpointed a single domestication event that happened in the Middle-Upper Amazon Basin in the mid-Holocene, mediated by indigenous people that had the forest as its source of livelihood long before European colonization. Finally, I compiled a biodiversity dataset of preserved specimen collections of *Theobroma*, and I present a by-product study that discussed the native area of *T. cacao* under a remote sensing approach. Results not only deeply impact our understanding on the evolution of wild cacao species, the flora and the people from Amazonia, but shed light on the study of the origin and history of wild crop relatives as a whole.

Keywords: Amazonia, domestication, natural selection, phylogenetics, populational genomics, taxonomy.

SUMMARY

1. INTRODUCTION **17**

2. PART I: TAXONOMIC TREATMENT **25**

2.1. CHAPTER 1. Unraveling cacaos: a taxonomic revision of *Theobroma* L. (Malvaceae, Byttnerioideae) **27**

2.2. CHAPTER 2. Three new species of wild cacaos (*Theobroma* sect. *Herrania*, Malvaceae) from the Western Amazon Basin **179**

2.3. CHAPTER 3. Proposal to conserve the name *Cacao sylvestris* (*Theobroma sylvestre*) (Malvaceae: Byttnerioideae) with a conserved type **199**

3. PART II: PHYLOGENETICS AND EVOLUTION **205**

3.1. CHAPTER 4. A framework for the study of the evolution of key crop traits in *Theobroma cacao* L. and its wild relatives **207**

3.2. CHAPTER 5. Phylogenetic evidence endorses the reestablishment of *Theobroma* including *Herrania* (Malvaceae, Byttnerioideae) **241**

4. PART III: POPULATION GENOMICS **259**

4.1. CHAPTER 6. Contrasting the effects of selection in allopatric and sympatric populations of cacao (*Theobroma cacao*) **261**

4.2. CHAPTER 7. Pre-Columbian domestication of *cupuaçu* (*Theobroma grandiflorum*), an Amazonian tree crop closely related to cacao **286**

5. PART IV: DATA MOBILIZATION AND BY-PRODUCT STUDIES **337**

5.1. CHAPTER 8. A taxonomic dataset of preserved specimen occurrences of *Theobroma* and *Herrania* (Malvaceae, Byttnerioideae) stored in 2020 ..... **339**

5.2. CHAPTER 9. Human influence on cacao (*Theobroma cacao* L.) dispersion as revealed by remote sensing data **373**

5. ELECTRONIC SUPPLEMENTARY DATASETS **399**

6. CONCLUSIONS **403**

INTRODUCTION

Tropical rainforest vegetations have always fascinated natural historians, from Humboldt, Martius, to Prance, Gentry and others (Teixeira, 1984; Miller *et al.*, 1996; Salgado, 2000; Peixoto and Morim, 2003; Pausas and Bond, 2018). However, we are still quite away from fully comprehending the complexities of the origin and evolution of particular groups of the Neotropical Region (Antonelli and Sanmartín, 2011; Antonelli *et al.*, 2018). After all, the Neotropics is the most diverse bioregion in terms of species richness and endemism in the globe, both in its fauna and flora (Hughes *et al.*, 2013; Ulloa-Ulloa *et al.*, 2017), whereas it is also one of the most threatened bioregions due to anthropogenic pressures (*e.g.*, Banda *et al.*, 2016).

This represents an enormous challenge for Brazil and South America in the study and preservation of its biota, as the continent bears the majority of plant species described for the world (Forzza *et al.*, 2012; Ulloa-Ulloa *et al.*, 2017). At the same time, the amount and availability of botanical records have significantly increased in the herbaria in the past few decades, so we now have access to vast herbarium collections going through a mass digitization framework (Pyke and Ehrich, 2014; Nualart *et al.*, 2017; Chapter 8). In Brazil, for example, innovative collaborative projects, namely the Brazilian Flora 2020 Project (BFG, 2018), are noteworthy, by inaugurating a novel era in the study of the Brazilian flora. This new framework allows new surveys with taxonomy, evolution, biogeography and conservation of particular genera or families (Wen *et al.*, 2015; Greve *et al.*, 2016; Nualart *et al.*, 2017). As a matter of fact, one of the targets of the Brazilian Flora 2020 Project – which is still under construction and requires further efforts from the taxonomist community – is to provide monographs with short descriptions for all algae, plants and fungi occurring in Brazil by at least 2020 (BFG, 2018).

Hence, studying the evolution and taxonomy of groups that have direct importance to society is the basis for development and innovation. Besides, this exalts our flora, as it also composes part of our history and cultural identity as a nation. In this project, we selected to focus on the “cacao group”, which encompasses flowering plants endemics to the Amazonian rainforests (Richardson *et al.*, 2015; Chapter 8) with species allocated in two genera: *Theobroma* L. and *Herrania* Goudot. Both are members from a representative and important tropical botanical family, Malvaceae, which has *ca.* 4,000 species in over 240 genera (Stevens, 2012). Malvaceae has an outstanding but still underassessed morphological and phylogenetic picture; it is so diverse that it is currently split into nine subfamilies based on molecular data (Alverson *et al.*, 1998; Baum *et al.*, 1998; Bayer *et al.*, 1999).

The genera *Theobroma* L. and *Herrania* Goudot (Malvaceae: Byttnerioideae)

Perhaps due to its long historical and economical importance, some *Theobroma* and *Herrania* species are very well known by many American societies. Nevertheless, although some species are quite known in applied research, it is shocking to realize that other wild cacao species remain unknown, with outdated information (Bletter and Daly, 2009). Two seminal taxonomic treatments for the cacao group are the revision of *Theobroma* (Cuatrecasas, 1964) and the synopsis of *Herrania* (Schultes, 1958). Both have provided one of the yet few broad attempts to properly describe the native species for each genus, recognizing 22 species for *Theobroma* and 17 for *Herrania* (Cuatrecasas, 1964; Schultes, 1958).

Supposedly, *Herrania* is distinguished from *Theobroma* by its compound leaves (*vs*. simple leaves in *Theobroma*), as well as by the trimerous calyx (*vs*. usually pentamerous in *Theobroma*) and for having the upper portion of an unguiculate petal much longer in *Herrania* than in *Theobroma* (Bletter and Daly, 2009). *Theobroma* was divided by Cuatrecasas (1964) into six sections: *T.* sect. *Andropetalum*, *T.* sect. *Glossopetalum*, *T.* sect. *Oreanthes*, *T.* sect. *Rhytidocarpus*, *T.* sect. *Telmatocarpus* and *T.* sect. *Theobroma*, based on morphological characters of branch growth, fruit, seeds, corolla and androecium. In *Herrania*, Schultes (1958) firstly recognized two sections – *H.* sect. *Herrania* and *H.* sect. *Subcymbicalyx* – based on the shape and size of the corolla and on the disposal and connation of the sepals. However, such characters, especially those related to the architecture of the branch growth and the seeds, are likely to be homoplastic (Cuatrecasas, 1964; Borrone *et al.*, 2007), which turns the circumscription of sections and subgroups (*e.g.*, subspecies and forms of *T. cacao*) problematic. Additionally, current questions regarding the taxonomy and evolution of the cacaos are still inconclusive, even after molecular phylogenetic analysis have been employed toward improvement of the systematics. Even basic aspects of pollination, dispersal modes and ecology of the groups remain poorly known in species with a narrower distribution. For instance, albeit Cuatrecasas (1964) suggests dispersal in *Theobroma* is mainly mediated by vertebrates, there are some notes of mammals, birds and water-mediated dispersal for other native species as well (Richardson *et al.*, 2015; Barbosa *et al.*, 2019).

Since the 1960s, many aspects involving the availability of botanical records and the structure of the International Code of Botanical Nomenclature have changed (Wen *et al.*, 2015; Nualart *et al.*, 2017), but the number of new described names in *Theobroma* and *Herrania* have not followed the same rate. Since the publication of the last revision of *Theobroma* (Cuatrecasas, 1964), we have spotted 5076 new records were collected and deposited at herbaria, which corresponds to 72% of all known records of the genus that are not revisited (Table 1). For *Herrania* (Schultes, 1958), this fraction reaches 63% (Table 1). Therefore, it is expected that undescribed taxa that occur in different parts of the Amazon might have been collected and deposited at herbaria, but these putatively new taxa were never analyzed to be formally described. Hence, the actual number of species in both genera is likely underestimated, as already pointed out by Cuatrecasas (1964).

In summary, there is a major demand for a cautious revision of collections from several Amazon herbaria, especially considering that this is a group whose material and symbolic importance are unquestionable. Such reevaluation is particularly relevant for the Brazilian collections, since they were not fully contemplated in the revision of Cuatrecasas (1964): important Amazonian collections deposited in INPA (Herbarium of the National Amazon Research Institute) and RB (Herbarium of the Rio de Janeiro Botanical Garden), currently reference collections of the Amazonian flora, were not mentioned by Cuatrecasas even after the publication of his monograph.

Advances with the advent of molecular systematics

Within the evolutionary scenario, it is long known that speciation is a key phenomenon that generates the observed species richness in the Neotropics and in South America. In 1859, Charles Darwin described his theory on the origin of species and explicitly expressed his desire to reconstruct the evolutionary history of species in the form of a “tree” that depicts species relationships. Today, we can assess this by using DNA sequencing data, and most phylogenies that have been generated under this approach are being used to enhance our understanding of speciation patterns and processes. However, changes at genomic level that cause speciation are still poorly understood in plants, and most of the taxonomic panorama for many genera and families are outdated or still underexplored.

Recent efforts have been undertaken to elucidate the phylogenetic history of native cacaos. Whitlock and Baum (1999) were the first to perform a broad-scale analysis using 11/22 species of *Theobroma* and 7/17 species of *Herrania* as ingroups. They used as markers *loci* from the vicilin gene that codes for a protein present in flowering plant seeds (Whiclock and Baum, 1999). The monophyly of both genera were recovered, but only *Herrania* was well-supported. Subsequently, Silva and Figueira (2005) made a second, independent analysis, based on another seed protein gene using, respectively, 11 and 3 *Theobroma* and *Herrania* terminals; the result, however, presented no significant differences in topology from the previous work. Borrone *et al.* (2007) also applied the same terminals, using five *loci* for three *WRKY* transcription factors. The phylogenetic inference presented a better performance when distinguishing different lineages of *Theobroma* and *Herrania*, but still was weakly supported in some branches. Some sections, such as *T.* sect *Glossopetalum*, emerged as paraphyletic, and the clade bearing species from *H.* sect. *Subcymbicalyx* consists of a major polytomy.

Finally, Richardson *et al.* (2015) published the first broad dated phylogeny for the Malvaceae using DNA sequences from molecular bank databases, where they reconstructed the phylogeny for the whole family, focusing on the lineages of *Theobroma* and *Herrania*. Essentially, they have incorporated *ndhF* chloroplast markers plus Borrone *et al.* (2007) data, solving some more relationships and recovering sister clades, but still not providing a fully solved tree for all infrageneric relationships.

Phylogenetic relationships between both genera, as well as within their respective sections, remain somewhat unclear, although already solved for some inner branches.

Given the published phylogenies so far, it is not possible yet to clearly establish the age of emblematic lineages such as of *T. cacao* (the commercial cacao) and *T. grandiflorum* (the "*cupuaçu*") (Whitlock and Baum, 1999; Richardson *et al.*, 2015). Hence, more than a species-level phylogenetic tree for the two genera – which is currently being generated by a post-doctoral researcher at the Rosario University, Dr. Ana Maria Bossa-Castro (J.E. Richardson, pers. comm.) –, novel population genomics approaches should also be used not only to properly assess these questions, but also to bring new knowledge on the evolution, diversification and speciation modes of the group.

Contemporary methods such as high-throughput DNA sequencing allow the simultaneous sequencing of different regions of the genome (Straub *et al.*, 2012; McKain *et al.*, 2018). This should provide significant information, especially in those occasions where the selection of fewer markers does not result in well supported phylogenetic resolution at infra-specific levels (Egan *et al.*, 2012; Twyford and Ennos, 2012). Genetic signatures of speciation modes can be detected when comparing genomes of sister-species that have likely evolved in sympatry or allopatry, as reported for animals (Noor, 1995; Coyne and Orr, 2004; Kang *et al.*, 2016). Such novel genomic methods are becoming increasingly less expensive and more accessible (Wicke and Schneeweiss, 2015), and, to our knowledge, a study exploring how changes at genomic level model different morphological structures in different speciation modes with a plant group would be pioneer. Moreover, due to its economic importance, many genomes of *T. cacao* have already been published, which facilitates the assembly of newly generated sequence data sets of its related species.

In summary, (i) the restricted Amazonian distribution of *Theobroma* and *Herrania* species; (ii) the relatively small number of species; (iii) the availability of taxonomic monographs and our proposed taxonomic treatment that outline key morphological differences amongst species; and (iv) the already sequenced; and (v) relatively small genome in *T. cacao* and *T. grandiflorum* make this an ideal group for studying speciation processes in tropical rainforest. This dissertation is organized in nine chapters and four parts: "taxonomic treatment", "phylogenetics and evolution", "populational genomics" and "data mobilization and by-product studies", each one tackling one or more goals associated to this project. The main goals of this dissertation are: (1) revisit the morphology and distribution of current taxa, reevaluating taxonomic delimitations at species levels; (2) look for morphological synapomorphies and discuss how the evolution of such characters might have occurred; (3) evaluate how genetic variation of populations of cacao (*T. cacao*) behave regarding its divergence levels of genes under selection; and (4) depict the origin and geographic history of *cupuaçu* (*T. grandiflorum*), a tree crop of relevance in Brazil, known for its humongous fruit and appreciated for its pulp.

References for this section

Alverson, W.S., Karol, K.G., Baum, D.A., Chase, M.W., Swensen, S.M., McCourt, R. and Sytsma, K. (1998) Circumscription of the Malvales and relationships to other Rosidae: evidence from *rbcL* sequence data. *American Journal of Botany*, **85**, 876-887.

Antonelli, A., Ariza, M., Albert, J., Andermann, T., Azevedo, J., Bacon, C., Faurby, S., … and Edwards, S.V. (2018) Conceptual and empirical advances in Neotropical biodiversity research. *PeerJ*, **6**, e5644.

Banda, R.K., Delgado-Salinas, A., Dexter, K.G., Linares-Palomino, R., Oliveira-Filho, A., Prado, D., Pullan, M., … and Pennington, R.T. (2016) Plant diversity patterns in neotropical dry forests and their conservation implications. *Science*, **353**, 1383–1387.

Barbosa, L., França, I. and Ruz, E.J.H. (2019) First records of *Theobroma speciosum* fruits dispersion. *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, **43**, 518-520.

Baum, D.A., Alverson, W.S. and Nyffeller, R. (1998) A durian by any other name: taxonomy and nomenclature of the core Malvales. *Harvard Papers in Botany*, **3**, 315-330.

Bayer, C., Fay, M.F., Brujin, A.Y., Savolainen, V., Morton, C.M., Kubitzki, J., Alverson, W.S. and Chase, M.W. (1999) Support for an expanded family concept of Malvaceae within a re-circumscribed order Malvales: combined analysis of plastid *atpB* and *rbcL* DNA sequences. *Botanical Journal of the Linnean Society*, **129**, 267-303.

Beaumont, M.A. (2005) Adaptation and speciation: what can Fst tell us? *Trends in Ecology and Evolution*, 20: 435-440.

BFG [The Brazil Flora Group] (2018) Brazilian Flora 2020: innovation and collaboration to meet Target 1 of the Global Strategy for Plant Conservation (GSPC). *Rodriguésia*, **69**, 1513-1527.

Bletter, N. and Daly, D.C. (2009) Cacao and its relatives in South America. In McNeil, C.L. (Ed.) *Chocolate in Mesoamerica: a cultural history of cacao*. Gainesville: University Press of Florida, pp. 31-68.

Borrone, J.W., Meerow, A.W., Kuhn, D.N., Whitlock, B.A. and Schnell, R.J. (2007) The potential of the WRKY gene family for phylogenetic reconstruction: an example from the Malvaceae. *Molecular Phylogenetics and Evolution*, **44**, 1141-1154.

Catchen, J., Hohenlohe, P., Bassham, S., Amores, A. and Creko, W. (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124-3140.

Coyne, J.A. and Orr, H.A. (2004) Speciation. Sunderland (MA): Sinauer Associates.

Cristóbal, C.L. (1976) Estudio taxonómico del género *Byttneria* (Sterculiaceae). *Bonplandia*, **4**, 1-428.

Cuatrecasas, J. (1964) Cacao and its allies: a taxonomic revision of the genus *Theobroma*. *Contributions from the United States National Herbarium*, **35**, 379-614.

DaCosta, J.M. and Sorenson, M.D. (2016) ddRAD-seq phylogenetics based on nucleotide, indel and presence-absence polymorphisms: analyses of two avian genera with contrasting histories. *Molecular Phylogenetics and Evolution*, **94**, 122-135.

Dorr, L.J. (1996) *Ayenia saligna* (Sterculiaceae), a new species from Colombia. *Brittonia*, **48**, 213-216.

Eaton, D.A.R. and Ree, R.H. (2013) Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology*, **62**, 689-706.

Egan, A.N., Schlueter, J. and Spooner, D.M. (2012) Applications of next-generation sequencing in plant biology. *American Journal of Botany*, **99**, 175-185.

Etter, P.D., Preston, J.L., Bassham, S., Cresko, W.A. and Johnson, E.A. (2011) Local De Novo assembly of RAD paired-end cotings using short sequencing reads. *PLOS ONE*, **6**, e18561.

Forzza, R.C., Baumgratz, J.F.A., Bicudo, C.E.M., Canhos, D.A.L., Carvalho Jr., A.A.C., Coelho, M.A.N., … and Zappi, D.C. (2012) New Brazilian floristic list highlights conservation challenges. *BioScience*, **62**, 39-45.

Greve, M., Lykke, A.M., Fagg, C.W., Gereau, R.E., Lewis, G.P., Marchant, R., Marshall, A.R., … and Svenning, J.C. (2016) Realising the potential of herbarium records for conservation biology. *South African Journal of Botany*, **105**, 317-323.

Hohenlohe, P.A., Bassham, S., Currey, M. and Cresko, W.A. (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Botanical Society London B*, **367**, 395-408.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. and Cresko, W.A. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS ONE Genetics*, **6**, e1000862.

Huey, R.B. and Pianka, E.R. (1977) Patterns of niche overlap among broadly sympatric versus narrowly sympatric Kalahari lizards (Scincidar: Mabuya). *Ecology*, 58: 119-128.

Hughes, C.E., Pennington, R.T. and Antonelli, A. (2013) Neotropical plant evolution: assembling the big picture. *Botanical Journal of the Linnean Society*, **171**, 1-18.

Jeffery, N.W., DiBacco, C., Wyngaarden, M.V., Hamilton, L.C., Stanley, R.R.E., Bernier, R., … and Bradbury, I.R. (2017) RAD sequencing reveals genomewide divergence between independent invasions of the European green crab *Carcinus maenas*) in the Northwestern Atlantic. *Ecology and Evolution*, 7**,** 2513-2524.

Kang, L., Settlage, R., McMahon, W., Michalak, K. Tae, H. Garner, H.R., Stacy, E.A., Price, D.K. and Michalak, P. (2016) Genomic signatures of speciation in sympatric and allopatric Hawaiian picture-winged *Drosophila*. *Genome Biology and Evolution*, **8** 1482-1488.

Kearse, M., Moir, R., Wilson, S., Stones-Havas, S., Cheung, M., Sturrock, S., … & Drummond, A. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28: 1647-1649.

Lawniczak, M.K.N., Emrich, S.J., Holloway, A.K., Regier, A.P., Olson, M., White, B., Redmond, S., Fulton, L. … and Besansky, N.J. (2010) Widespread divergence between incipient *Anopheles* gambiae species revealed by whole genome sequences. *Science*, **330**, 512-514.

Maddison, W.P. and Maddison, D.R. (2018) Mesquite: a modular system for evolutionary analysis. Version 3.51. http://www.mesquiteproject.org.

Mattos, J.R., Ferreira, C.D.M., Bovini, M.G. and Coelho, M.A.N. (2019) Malvaceae cultivada no Arboreto do Jardim Botânico do Rio de Janeiro: a família dos hibiscos (1 ed.) Rio de Janeiro: Vertente edições. 136p.

McKain, M.R., Johnson, M.G., Uribe-Conversa, S., Eaton, D. and Yang, Y. (2018) Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, **6**, e1038.

Miller, J.S., Barkley, T.M., Iltis, H.H., Lewis, W.H., Forero, E., Plotkin, M., Phillips, O., Rueda, R. and Raven, P.H. (1996) Alwyn Howard Gentry, 1945-1993: a tribute. *Annals of the Missouri Botanical Garden*, **4**, 433-460.

Morrone, J.J. (2014) Biogeographical regionalisation of the Neotropical region. *Zootaxa*, **3782**, 1-110.

Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., … and Jiggins, C.D. (2013) Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology*, **22**, 814-826.

Noor, M.A. (1995) Speciation driven by natural selection in *Drosophila*. *Nature,* **375**, 674-675.

Nualart, N., Ibañez, N., Soriano, I. and López-Pujol, J. (2017) Assessing the relevance of herbarium collections as tools for conservation biology. *Botanical Revision*, **83**, 303-325.

Olson, D.M., Dinerstein, E., Wikramanayake, R.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., … and Kassem, K.R. (2001) Terrestrial ecoregions of the world: a new map of life on earth:

a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, **51**, 933-938.

Pausas, J.G. and Bond, W.J. (2018) Humboldt and the reinvention of nature. *Journal of Ecology*, *in press.*

Peixoto, A.L. and Morim, M.P. (2003) Coleções botânicas: documentação da biodiversidade brasileira. *Ciência e Cultura*, **55**, 21-24.

Pimentel, D., Smith, G.J.C. and Soans, J. (1967) A population model of sympatric speciation. *The American Naturalist*, 101: 493-504.

Pyke, G.H. and Ehrlich, P.R. (2010) Biological collections and ecological/environmental research: a revision, some observations and a look to the future. *Biological Revisions*, **85**, 247-266.

Revell, L.J. (2011) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution,* 3, 217-223.

Richardson, J.E., Whitlock, B.A., Meerow, A.W. and Madriñan, S. (2015) The age of chocolate: a diversification history of *Theobroma* and Malvaceae. *Frontiers in Ecology and Evolution*, **3**, a120.

Rousset, F. (2008) Genepop'007: a complete reimplementation of the Gene-pop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103-106.

Salgado, G.M.L. (2000) História e natureza em von Martius: esquadrinhando o Brasil para construir a nação. *História, Ciências, Saúde*, **7**, 391-413.

Schultes, R.E. (1958) A synopsis of the genus *Herrania*. *Journal of the Arnold Arboretum*, **34**, 217-278.

Silva, C.R.S. and Figueira, A. (2005) Phylogenetic analyses of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences. *Plant Systematics and Evolution*, **250**, 93-104.

Silva, C.R.S., Venturieri, G.A. and Figueira, A. (2004) Description of Amazonian *Theobroma* L. collections, species identification, and characterization of interspecific hybrids. *Acta Botanica Brasilica*, **18**, 333-341.

Soans, A.B., Pimentel, D. amd Soans, J.S. (1974) Evolution of reproductive isolation in allopatric and sympatric populations. *The American Naturalist*, 108, 117-124.

Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C. and Liston, A. (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349-364.

Tajima F 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Teixeira, A.R. (1984) O "Programa Flora" do Brasil - história e situação atual. *Acta Amazonica*, **14**, 31-47.

Thiers, B. [continuously updated]. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/.

Twyford, A.D. and Ennos, R.A. (2012) Next-generation hybridization and introgression. *Heredity*, **108**, 179-189.

Ulloa-Ulloa, C.U., Acevedo-Rodríguez, P., Beck, S., Belgrano, M.J., Bernal, R., Berry, P.E., Brako, L., ... and Jørgensen, P.M. (2017) An integrated assessment of the vascular plant species of the Americas. *Science*, **358**, 1614-1617.

Weir, B.S. and Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 358-1370.

Wen, J., Ickert-Bond, S.M., Appelhans, M.S., Dorr, L.J. and Funk, V.A. (2015) Collections-based systematics: opportunities and outlook for 2050. *Journal of Systematics and Evolution*, **53**, 477-488.

Whitlock, B.A. and Baum, D.A. (1999) Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene *vicilin*. *Systematic Botany*, **24**, 128-138.

CONCLUSIONS

The title of this dissertation is “Unraveling cacaos: systematics and evolution of *Theobroma* L. and *Herrania* Goudot (Malvaceae, Byttnerioideae)”. After four parts and nine chapters, I can say the we indeed “unraveled” cacaos towards the understanding of its systematics, evolution and genomics. The main “take home messages” that can be pinpointed from this dissertation are:

1. There are 35 wild cacao species known on Earth. Our taxonomic revision (Chapter 1), in light of phylogenetic evidence (Chapter 4) recognizes 35 species of *Theobroma* allocated into six sections, differentiated mostly by floral and fruit features. This number already considers three new species of cacaos (*T. globosum*, *T. nervosum* and *T. schultesii*) newly described in here (Chapter 2).

2. *Theobroma* is not monophyletic. Phylogenetic evidence (Chapter 4) suggested that the genus *Theobroma* is not monophyletic, with a closely related genus, *Herrania* nested in it. This was the basis for me to provide the recircumscription of *Theobroma sensu lato*, including *Herrania* as a section of the first (Chapter 5).

3. The history of wild cacao species is deeply marked by human domestication. Based on genomic Chapter 6, Chapter 7) and remote sensing evidence (Chapter 9), we have shown that the origin and geographic history of a number of species wild cacaos is deeply marked by human domestication, which have selected particular features towards obtaining desirable features, mostly for the use of the fruit pulp (in the case of *cupuaçu*), or the seeds (especially for the chocolate industry). Other species of cacao are likely to have been human-dispersed or manipulated to some degree as well. Additionally, by using wild cacao species as study-case, we could demonstrate how many ancient societies that lived in Amazonia long before European colonization have managed the forest in a sustainable way, as evidenced by the creation of *cupuaçu* (Chapter 7) already five to seven millennia ago.

Much has yet to me made with the study of cacao and relatives, both in terms of understanding more with detail the diversification and differences of particular clades (such as section *Herrania*), or to explore the origin of other species that likely were domesticated (such as *T. bicolor*). It is a life-time work to be made not only in the field of systematics and evolution of the Malvaceae, but also in the field of biotechnology, crop genomics and even archeology and anthropology.