UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO DE PLANTAS

ESTADO ATUAL DO CONHECIMENTO CIENTÍFICO SOBRE CARYOCARACEAE E RECURSOS GENÔMICOS PARA O PEQUIZEIRO (*Caryocar brasiliense* Camb.)

RHEWTER NUNES

Orientadora:
Prof.ª Mariana Pires de Campos Telles

Outubro – 2019

sistema de bibliotecas ufg

PRPG
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[ ] Dissertação** **[x] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Rhewter Nunes

Título do trabalho: Estado atual do conhecimento científico sobre Caryocaraceae e recursos genômicos para o pequizeiro (*Caryocar brasiliense* camb.).

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [ ] SIM [x] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Rhewter Nunes
Assinatura do(a) autor(a)[2]

Ciente e de acordo:

Assinatura do(a) orientador(a)[2]

Data: 16 / 01 / 2020

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

[2] A assinatura deve ser escaneada.

RHEWTER NUNES

ESTADO ATUAL DO CONHECIMENTO CIENTÍFICO SOBRE CARYOCARACEAE E RECURSOS GENÔMICOS PARA O PEQUIZEIRO (*Caryocar brasiliense* Camb.)

Tese apresentada ao programa de pós-graduação em Genética e Melhoramento de Plantas, da Universidade Federal de Goiás, como requisito parcial para a obtenção do título de Doutor em Genética e Melhoramento de Plantas.

Orientadora:

Prof.ª Dr.ª Mariana Pires de Campos Telles

Goiânia, GO – Brasil

2019

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Nunes, Rhewter
ESTADO ATUAL DO CONHECIMENTO CIENTÍFICO SOBRE CARYOCARACEAE E RECURSOS GENÔMICOS PARA O PEQUIZEIRO (Caryocar brasiliense Camb.) [manuscrito] / Rhewter Nunes. - 2019.
104 f.: il.

Orientador: Profa. Dra. Mariana Pires de Campos Telles.
Tese (Doutorado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-graduação em Genética e Melhoramento de Plantas, Goiânia, 2019.
Bibliografia. Apêndice.

1. Cerrado. 2. genômica comparativa. 3. recursos genéticos. 4. SSR. 5. pequi. I. Telles, Mariana Pires de Campos, orient. II. Título.

CDU 575

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE AGRONOMIA

ATA DE DEFESA DE TESE

Ata Nº **007/2019** da sessão de Defesa de Tese de **Rhewter Nunes** que confere o título de Doutor(a) em **Genética e Melhoramento de Plantas**, na área de concentração em **Genética e Melhoramento de Plantas.**

Ao/s **Trinta e um dias de outubro de dois mil e dezenove**, a partir da(s) **13:30hrs**, no(a) **Auditório** do **ICB 5**, realizou-se a sessão pública de Defesa de Tese intitulada "**ESTADO ATUAL DO CONHECIMENTO CIENTÍFICO SOBRE CARYOCARACEAE E RECURSOS GENÔMICOS PARA O PEQUIZEIRO (Caryocar brasiliense Camb.)**". Os trabalhos foram instalados pelo(a) Orientador(a), Professor(a) Doutor(a) **Mariana Pires de Campos Telles ICB/UFG** com a participação dos demais membros da Banca Examinadora: Professor(a) Doutor(a) **Flávia Melo Rodrigues UEG/GO**, membro titular externo; Professor(a) Doutor(a) **Rosana Pereira Vianello - EMBRAPA**, membro titular externo; Professor(a) Doutor(a) **José Alexandre Felizola Diniz Filho / UFG**, membro titular interno; Professor(a) Doutor(a) **Thannya Nascimento Soares ICB/UFG**, membro titular interno. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do **trabalho**. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Tese tendo sido(a) o(a) candidato(a) **aprovado(a)** pelos seus membros. Proclamados os resultados pelo(a) Professor(a) Doutor(a) **Mariana Pires de Campos Telles ICB/UFG**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, ao(s) **Trinta e um dias de outubro de dois e dezenove**.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Mariana Pires De Campos Telles**, **Presidenta**, em 18/11/2019, às 10:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **José Alexandre Felizola Diniz Filho**, **Professor do Magistério Superior**, em 18/11/2019, às 10:38, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Thannya Nascimento Soares**, **Coordenadora de Curso**, em 18/11/2019, às 10:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **RHEWTER NUNES**, **Discente**, em

18/11/2019, às 11:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rosana Pereira Vianello**, **Usuário Externo**, em 18/11/2019, às 12:28, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Flávia Melo Rodrigues**, **Usuário Externo**, em 18/11/2019, às 14:37, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1009032** e o código CRC **D0DC0096**.

Referência: Processo nº 23070.036959/2019-73 SEI nº 1009032

A tod@s aquel@s que, por conta de sua classe social, gênero, cor ou sexualidade, não tiveram a possibilidade de se apaixonar pela ciência, ofereço.

A minha mãe (Lucia) e minha avó (Maria Olinda) por todo cuidado, incentivo e torcida, dedico.

AGRADECIMENTOS

Essa tese foi desenvolvida no contexto do grupo de trabalho de Genética e Genômica Evolutiva do Instituto Nacional de Ciência e Tecnologia – Ecologia, Evolução e Conservação da Biodiversidade (INCT - EECBio), financiado pelo CNPq (Processo nº 402178/2016-5) e Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG – Chamada nº 07/2014, Processo: nº 201410267001736). Além disso, contou com o financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo projeto universal chamada (processo 435477/2018-8). Durante os primeiros anos de doutorado, recebi uma bolsa (Demanda Social) da Coordenação de Pessoal de Nível Superior (CAPES). Agradeço também à equipe do INCT em Ecologia, Evolução e Conservação da Biodiversidade (EECBio) pela oportunidade de uma bolsa DTI que permitiu a finalização do doutorado. Esse agradecimento se estende então aos professores José Alexandre Diniz Filho, Mariana Telles e Thiago Rangel. Nesse sentido, agradeço a todos os órgãos de fomento que possibilitaram que esse trabalho fosse realizado.

Gostaria de agradecer à Universidade Federal de Goiás e ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas pela estrutura e possibilidade de concluir esse trabalho. Agradeço também à todos os professores do programa que compartilharam comigo seus conhecimentos e contribuíram para a minha formação pessoal e profissional. Além destes, também agradeço aos demais professores que fizeram parte da minha formação.

Agradeço a minha orientadora e amiga, Mariana Telles, por ser a melhor mãe científica que eu poderia ter. Agradeço o compartilhamento de vivências, o acolhimento, a paciência e o suporte. Obrigado por ter feito a diferença na minha vida e por sempre me inspirar a ser uma pessoa melhor e a amar o meu trabalho. Obrigado por todas as oportunidades, por acreditar na qualidade meu trabalho (muitas vezes em momentos que eu mesmo duvidava) e por sempre me incentivar a seguir em frente e a fazer as coisas acontecerem. O carinho, a admiração e gratidão que tenho pela senhora se transpõem ao que pode ser escrito em um breve parágrafo de agradecimentos.

Agradeço à todos os membros do Laboratório de Genética & Biodiversidade pela possibilidade de trabalho em conjunto e aprendizado. Agradeço às professoras Mariana, Thannya e Rosane pela administração e manutenção desse laboratório onde pude desenvolver minhas atividades. Agradeço aos meus colegas de laboratório Natácia, Adriana, Vanessa, Kassia, Rejane, Edivaldo, Ueric, Andrezza, Tatiane, Anita e Rodrigo pela troca de vivências, companhia e conhecimento. Em especial, gostaria de agradecer à Ramilla, à Cintia, à Ariany e à Fernanda, por terem me acolhido quando cheguei ao LGBio e por terem me feito sentir que ali era a minha nova casa.

Gostaria também de fazer um agradecimento especial aos estudantes que eu tive a oportunidade de auxiliar seja na forma de co-orientação ou como colaborador em seus trabalhos de conclusão de curso, iniciação científica ou mestrado. Agradeço então: Amanda Melo, Amanda Apolinário, Lais, Leonardo, Larissa Matos, Larissa Ferreira, Larissa Carvalho, Maria Eduarda e Sara. As relações que consegui construir com vocês me mostraram que eu estou no caminho certo e que o processo de orientação, ainda que difícil, é uma das atividades que mais me trazem satisfação. Foi muito bom auxiliar vocês a escrevem seus primeiros textos científicos, a montarem seus pôsteres para congresso, ver vocês ganhando premiações com os nossos trabalhos em eventos científicos, sendo aprovados em programas de mestrado e dentre todas as demais coisas que demonstram o crescimento profissional e pessoal de vocês. Tenho muito orgulho de fazer parte da história de vocês.

Em aspectos da minha formação profissional e pessoal eu também preciso agradecer aos membros do cursinho PreparaTrans. Esse é um cursinho pré-vestibular voltado principalmente para pessoas transexuais e outros LGBTs mas que também atende pessoas de outros grupos em vulnerabilidade social e que é sediado na Faculdade de Educação da UFG. Eu tive a oportunidade de trabalhar voluntariamente nesse projeto ministrando au de biologia e matemática e foi uma experiência transformadora para mim. Eu pude perce o quão importante é não desassociar as minhas práticas enquanto pesquisador das pautas cunho social ainda mais em uma país com características como as do nosso. Muito obrigado por todo aprendizado.

Gostaria de agradecer a todas as pessoas que formaram a minha rede de apoio emocional durante o período do doutorado. Agradeço a minha família, em especial minha mãe Lucia e minha avó Maria Olinda por entenderem minha ausência e por torcerem por

mim. A minha avó agradeço por ter sempre cuidado de mim e por ser a principal responsável pela minha formação moral. Te amo. Agradeço também meu padrasto Manuel pelo apoio e torcida. Agradeço também a minha segunda família: Alexandre, Alessandra, Luana e Noerci por todo carinho e apoio dedicados a mim. Agradeço aos meus filhos felinos Vincent (Vinvin), Tarsila (Neném) e Magali (Mag). Amo vocês. Agradeço à Ivone, à Stela e à Isabela por terem sido as minhas primeiras amigas em Goiânia e por terem compartilhado tantos momentos importantes comigo. Vocês são muito importantes para mim. Agradeço aos amigos de fora da academia que tiveram muita paciência comigo, sobretudo no período de conclusão da tese: Julio, Thamirys e Giorgia. Agradeço à Suzy e ao Guilherme por terem me aguentado no tempo que moramos juntos e pela amizade ao dividirmos a vida nesse período. Também agradeço especialmente aos meus amigos Mari e Lucas, por serem o meu casal favorito e por nunca terem me desamparado. Muito da conclusão desse doutorado tem a contribuição de vocês. Também agradeço aos muitos amigos da ecologia que fiz ao vir trabalhar no INCT tais como Raísa, Milena, Elisa, Lucas Jardim, Bruno, Kelly, Priscila, Daisy e, em especial, Laura por todo companheirismo e por ter se tornado uma amiga que quero levar por toda a vida. Agradeço também aos amigos mais recentes trazidos pelo Alexandre e que foram sempre acolhedores comigo: Gabriella, João Vitor, Thayná, Geovane, Débora, Fernanda, Gabriel, Lucas, Janaina, Luiza, Aryadine, Samanta, Stefany e Stéphane.

Agradeço também a quem por ventura eu tenha esquecido de mencionar aqui. Muito obrigado a todos que contribuíram para que esse trabalho existisse.

SUMÁRIO

NUNES, R. **ESTADO ATUAL DO CONHECIMENTO CIENTÍFICO SOBRE CARYOCARACEAE E RECURSOS GENÔMICOS PARA O PEQUIZEIRO (*Caryocar brasiliense* Camb.).** 2019. 104 f. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2019.[1]

O desenvolvimento de novas tecnologias de sequenciamento de DNA, acompanhado pela redução de custos para a obtenção desses dados tem melhorado o acesso e permitido a caracterização dos recursos genômicos de espécies vegetais. No entanto, as espécies vegetais silvestres, que se configuram como recursos genéticos, têm sido negligenciadas e subutilizadas, com poucas informações em nível genômico. A geração de recursos genômicos para essas espécies pode aumentar as informações sobre aspectos biológicos importantes para contribuir com estudos evolutivos, de domesticação e em processos de utilização, além de possibilitar sua popularização e conservação. Nesse contexto, nesse trabalho foi realizado um levantamento do conhecimento científico sobre as espécies da família de plantas Neotropicais Caryocaraceae e foram disponibilizados recursos genômicos em larga escala para *Caryocar brasiliense* (Pequizeiro), um importante recurso genético do Cerrado brasileiro. Utilizando dados de sequenciamento de alto desempenho, foi realizada a montagem do rascunho do genoma nuclear de *C. brasiliense* e a montagem do genoma completo do cloroplasto. O conhecimento científico sobre a família Caryocaraceae é centrado em algumas poucas espécies do gênero Caryocar, especialmente *C. brasiliense*. Além disso, a maioria dos trabalhos foram realizados por pesquisadores brasileiros. Quanto aos recursos genômicos de *C. brasiliense*, o rascunho do genoma apresentou um tamanho total de 464.365.380 pb, sendo o maior *contig* com o tamanho de 64.707 pb e metade dos *contigs* contidos em sequências de pelo menos 6005 pb. Essa montagem equivale a 45,69 % da estimativa do tamanho do genoma nuclear, obtida com base na distribuição de frequência de *k-mers*. O rascunho do genoma foi utilizado no desenho de 30 pares de iniciadores para amplificação de regiões microssatélites divididos em cinco conjuntos de multiplex de PCR. Quanto ao genoma do cloroplasto, foi observado um tamanho de 165.793 bp e recuperado em um único cromossomo circular. O genoma do cloroplasto de *C. brasiliense* apresenta uma estrutura quadripartida, comumente encontrada em outras angiospermas, com a presenta de uma região longa única de 84.137 bp, uma região curta única de 11.852 bp e um par de regiões invertidas duplicadas de 34.902 bp cada. As análises de genômica comparativa sugerem uma expansão das regiões invertidas repetidas no pequizeiro quando comparado com outras espécies da ordem Malpighiales. Além disso, a análise filogenética fornece evidências de que Caryocaraceae se apresenta como um grupo irmão dos demais membros da ordem Malpighiales, com altos valores de suporte para os nós. Esse trabalho fornece os primeiros recursos genômicos em larga escala para uma espécie da família Caryocaraceae. O conjunto de iniciadores aqui descrito pode ser utilizado no desenvolvimento de novos marcadores moleculares para *C. brasiliense* e a sequência do genoma do cloroplasto fornece uma melhor compreensão sobre as relações evolutivas dentro da ordem Malpighiales.

Palavras-chave: Cerrado, genômica comparativa, recursos genéticos, SSR.

[1] Orientadora: Profª. Drª. Mariana Pires de Campos Telles. ICB – UFG.

NUNES, R. **CURRENT STATE OF SCIENTIFIC KNOWLEDGE ON CARYOCARACEAE AND GENOMIC RESOURCES FOR THE PEQUIZEIRO (Caryocar brasiliense Camb.).** 2019. 104 f. Thesis (PhD in Genetics and Plant Breeding) - School of Agronomy, Federal University of Goiás, Goiânia, 2019.[1]

The development of new DNA sequencing technologies, accompanied by reduced costs to obtain these data has improved access and allowed the characterization of genomic resources of plant species. However, wild plant species, which are configured as genetic resources, have been neglected and underused, with little information at the genomic level. The generation of genomic resources for these species can increase the information on important biological aspects to contribute to evolutionary studies, use and domestication processes, as well as enabling their popularization and conservation. In this context, in this work a survey of scientific knowledge about the species of the Neotropical plant family, Caryocaraceae, was carried out and large-scale genomic resources were made available for *Caryocar brasiliense* (Pequizeiro), an important genetic resource of the Brazilian Cerrado. Using high throughput sequencing data, a draft of the *C. brasiliense* nuclear genome was assembled and the complete chloroplast genome was assembled. Scientific knowledge about the Caryocaraceae family is centered on a few *Caryocar* genera species, especially *C. brasiliense*. In addition, most of the work was done by Brazilian researchers. Regarding the 0genomic resources of *C. brasiliense*, the genome draft had a total size of 464,365,380 bp, the largest contig having a size of 64,707 bp and half of the contigs contained in sequences of at least 6005 bp. This assembly is equivalent to 45.69% of the estimated nuclear genome size obtained from the frequency distribution of k-mers. The draft genome was used in the design of 30 primer pairs for amplification of microsatellite regions divided into five PCR multiplex sets. As for the chloroplast genome, a size of 165,793 bp was observed and recovered in a single circular chromosome. The *C. brasiliense* chloroplast genome has a quadripartite structure, commonly found in other angiosperms, with a single long region of 84,137 bp, a single short region of 11,852 bp and a pair of duplicated inverted regions of 34,902 bp each. Comparative genomics analyzes suggest an expansion of the inverted inverted regions in pequi tree when compared to other species of the order Malpighiales. In addition, phylogenetic analysis provides evidence that Caryocaraceae presents itself as a sister group of the other members of the order Malpighiales, with high support values for the nodes. This work provides the first large-scale genomic resources for a species in the Caryocaraceae family. The set of primers described herein can be used to develop new molecular markers for *C. brasiliense* and the chloroplast genome sequence provides a better understanding of the evolutionary relationships within the Malpighiales order.

Key words: Cerrado, genetic resources, comparative genomics, SSR.

---

[1] Advisor: Dr. Mariana Pires de Campos Telles. ICB – UFG.

1 INTRODUÇÃO GERAL

A obtenção de dados biológicos, especialmente dos dados de natureza molecular, se desenvolveu rapidamente desde a segunda metade do século XX (METZKER, 2005). Isso se deve principalmente pela ocorrência de diversos eventos de inovação tecnológica que geraram mudanças de paradigma no que se refere ao aumento do volume de dados obtidos (METZKER, 2010). Essa possibilidade de obtenção de dados em larga escala culminou no desenvolvimento de novos campos nas Ciências Biológicas conhecidos como "Ciências Ômicas". Esse termo tem origem do neologismo da língua inglesa "omics" e se refere a campos tais como a Genômica, a Transcritômica, a Proteômica, a Metabolômica e a Fenômica, relacionando-se à obtenção e análise de dados em larga escala de genomas, transcritos, proteínas, metabólitos e fenótipos, respectivamente (CARVALHO et al., 2019; GOMEZ-CABRERO et al., 2014; WARD; WHITE, 2002).

Em se tratando da Genômica, por exemplo, partiu-se do sequenciamento de 5.375 pares de base do genoma do bacteriófago ΦX174 no ano de 1977 (SANGER et al., 1977) para cerca de 0,1% (~7.700.000) da população mundial de humanos com seus genomas sequenciados em algum nível em 2017 (SHENDURE et al., 2017). Os avanços nesse campo estão relacionados principalmente ao desenvolvimento de novas tecnologias de sequenciamento de DNA e dos métodos e ferramentas de análise de dados. Um dos principais marcos na Genômica foi o Projeto Genoma Humano (PGH), uma iniciativa internacional que envolveu milhares de pesquisadores e milhões de dólares com o objetivo de obter as sequências completas de nucleotídeos que compõem os cromossomos humanos (COLLINS, 2003; REISS, 2001). O PGH alcançou um grande apelo popular, principalmente pela expectativa da possibilidade de se desvendar a origem de várias doenças que acometem a espécie humana.

Com o PGH e outros projetos genoma para outras espécies, surgiu a necessidade de crescimento de um campo de interface entre a Ciência da Computação, a Estatística e a Genética: a Bioinformática. Esse campo está relacionado a geração, armazenamento e

análise de dados moleculares em larga escala (OUZOUNIS, 2012). Ainda que esse campo tenha surgido na década de 1960 com a conversão de códigos de aminoácidos de três letras para uma única letra (e o consequente ganho de otimização de espaço de armazenamento de sequências), a sua consolidação se deu pela necessidade de se decifrar a complexidade presente das sequências obtidas em projetos genoma (JOSEPH; NAIR, 2012).

Em proximidade histórica ao desenvolvimento do Projeto Genoma Humano, outros projetos genoma também foram desenvolvidos (MICHAEL; JACKSON, 2013; TÜRKTAŞ; KURTOĞLU; DORADO, 2015). Para a genômica vegetal, alguns dos marcos científico-históricos mais importantes que podem ser utilizados como exemplo são: a publicação do genoma de *Arabidopsis thaliana*, uma importante planta modelo, no ano 2000 (KAUL et al., 2000); as publicações dos genomas de duas subespécies de arroz (*Oryza sativa* ssp. indica e *Oryza sativa* ssp. japonica), os primeiros genomas de uma planta cultivada e de importância econômica, em 2002 (GOFF, 2002; YU, 2002); a publicação do primeiro genoma de uma espécie de alga (*Cyanidioschyzon merolae*), em 2004 (TANAKA et al., 2004); e a publicação do primeiro genoma de uma espécie arbórea (*Populus trichocarpa*), em 2006 (TUSKAN et al., 2006).

Esses marcos histórico-científicos de sequenciamento de genomas de plantas revolucionaram a compreensão da biologia das espécies alvo e possibilitaram/possibilitam o desenvolvimento de diversos outros estudos com espécies relacionadas (LEITCH; BOTANIC; MARY, 2017). Para espécies como *Oryza sativa*, por exemplo, ocorreu um incremento de duas vezes o número de artigos científicos após a publicação dos genomas das duas subespécies segundo um estudo cientométrico com todos os trabalhos com essa espécie entre os anos de 1985 e 2014 (LIU; ZHANG; WANG, 2017). Esse crescimento provavelmente está relacionado com a disponibilidade de recursos genômicos e a possibilidade de diferentes abordagens de estudo para a espécie que possui um genoma sequenciado (LIU; ZHANG; WANG, 2017).

Atualmente (junho de 2019), 380 espécies do grupo Embryophyta possuem um genoma de referência depositado no *National Center for Biotechnology Information* (NCBI). Também foram depositados, até então, 2638 genomas de cloroplasto e 186 genomas de mitocôndrias de plantas. Considerando os genomas nucleares, 85,52% (n= 325) deles foram sequenciados nos últimos seis anos demonstrando tanto o desenvolvimento das tecnologias de sequenciamento de DNA quanto o aumento no interesse de se gerar esse tipo de recurso

genômico. Além disso, iniciativas internacionais que visam expandir o número de espécies com genoma sequenciado tem sido criadas na comunidade científica (CHENG et al., 2018; LEWIN et al., 2018).

Segundo dados do *Royal Botanic Gardens Kew*, 57,7% das plantas que tiveram seus genomas sequenciados até o momento são espécies cultivadas, seguidas por 22,3% de espécies modelo ou relacionadas com as mesmas, 17,7% de espécies relacionadas com as espécies cultivadas e apenas 2,3% de outros tipos de espécies de plantas. Dentre as espécies cultivadas, 37,7% delas são de plantas utilizadas para alimentação humana, 7,3% de plantas importantes para a indústria de matérias, 4,1% de plantas de importância medicinal, 2,7% para forragem de animais domésticos, 2,3% para indústria de combustíveis e 3,6% de outras plantas cultivadas. Nesse sentido, a maioria dos esforços para se obter sequências de genomas de plantas tem sido focada em espécies que já são utilizadas para consumo humano (LEITCH; BOTANIC; MARY, 2017; MICHAEL; JACKSON, 2013).

Com a redução do custo das tecnologias de sequenciamento de DNA, a possibilidade de expandir a obtenção desse tipo de dado para espécies negligenciadas ou subutilizadas, mas que se configuram como importantes recursos genéticos pelo seu potencial de uso, tem se tornado cada vez mais possível (DAWSON et al., 2009; MAYES et al., 2012). A obtenção de recursos genômicos para espécies com esse perfil pode auxiliar em um melhor conhecimento das potencialidades das mesmas, bem como, em seus processos de domesticação, manejo e conservação (ANGELONI et al., 2011; GARNER et al., 2016; MANEL et al., 2016). A criação de projetos genoma pode auxiliar no processo de agregação de valor para espécies negligenciadas e/ou subutilizadas e favorecer o desenvolvimento de estratégias de conservação de suas populações nativas.

Nesse contexto, o Cerrado brasileiro é rico em espécies negligenciadas ou subutilizadas e que são recursos genéticos para alimentação humana e animal, fabricação de materiais como madeira e látex, ornamentação e fármacos (TUNHOLI; RAMOS; SCARIOT, 2013). Como exemplos pode-se citar algumas consideradas prioritárias para a região centro oeste, tais como: pequi (*Caryocar brasiliense*), cagaita (*Eugenia dysenterica*), baru (*Dipteryx alata*), araticum (*Anona crassiflora*), mangaba (*Hancornia speciosa*), dentre outras. Dentre as espécies nativas do Cerrado, uma das mais populares e com maior apelo para incorporação em sistemas tradicionais de cultivo é o pequi/pequizeiro (*C. brasiliense* Camb.) (ARAUJO, 1995; VIEIRA et al., 2016). Essa espécie é amplamente conhecida no

Brasil como um símbolo do estado de Goiás e é utilizado na culinária tradicional. Seus frutos são comumente adquiridos pelas populações humanas locais por extrativismo e são comercializados em beiras de rodovias, feiras e até mesmo mercados.

Diversos fatores reforçam a importância de *C. brasiliense* enquanto recurso genético do Cerrado. Existem atualmente (junho de 2019) oito patentes registradas com produtos desenvolvidos a partir do pequi (102017027538, 102014024970, 2014131059, 102013020796, 102013212619, PI0601631, PI0206337 e 2001064145). Três dessas patentes foram concedidas a pessoas e instituições com sede fora do Brasil e se referem a utilização do pequi na forma de óleo no tratamento capilar; como um dos componentes de um extrato profilático para hipofunção biológica causada por estresse ambiental ou envelhecimento de pele, cabelo e cavidade oral; e um método de obtenção de lipídeos. As demais patentes são relacionadas a capsulas como suplemento vitamínico, antioxidante e antimutagênico, um tablete comestível de pequi, um processo de obtenção de pectina do pericarpo dos frutos de pequi, o uso do óleo de pequi na flotação de minerais (patente da Universidade Federal de Goiás) e do uso do carvão obtido das cascas de pequi no tratamento de água contaminada com glifosato (inpi, 2019).

Além disso, diversos estudos têm apontado para o alto valor nutricional dos frutos de pequi e das suas propriedades antioxidantes e antimutagênicas. A polpa do fruto do pequizeiro é rica em ácidos graxos insaturados, vitaminas e ácidos fenólicos, além de carotenóides, como violaxantina, luteína e zeaxantina (MARIANO; COURI; FREITAS, 2009; PIANOVSKI et al., 2008). Também foram observados efeitos benéficos em inflamações geradas por atividade física em atletas que praticam esportes de longa duração e atividade intensa (MIRANDA-VILELA et al., 2009; ROLL et al., 2018), além de reduzir significativamente o estresse oxidativo induzido pelo uretano, protegendo contra a genotoxicidade *in vivo* (COLOMBO et al., 2015).

Considerando a grande importância do pequizeiro e a necessidade de se gerar recursos genômicos para espécies com esse perfil, nesse trabalho foram disponibilizados recursos genômicos em larga escala para *C. brasiliense*. Aqui, considera-se como recurso genômico todo tipo de informação oriunda do genoma de uma espécie e que pode ser utilizada como informação ou ferramenta molecular no estudo da biologia de populações e indivíduos da espécie em questão (Ex: genes, elementos transponíveis, microssatélites, sequências completas ou parciais do genoma, etc).

Nesse contexto, essa tese foi subdividida em três capítulos principais. No primeiro capítulo foi realizado um levantamento do atual estado de conhecimento científico sobre a família Caryocaraceae. O segundo capítulo disponibilizou um rascunho do genoma nuclear de *C. brasiliense* e um conjunto de iniciadores para amplificação de regiões microssatélites em multiplex, disponibilizando novos marcadores moleculares para essa espécie. No terceiro e último capítulo, foi realizada a montagem completa do genoma do cloroplasto de *C. brasiliense*, utilizando uma abordagem de genômica comparativa para auxiliar na compreensão da evolução de genomas de cloroplasto de espécies da ordem Malpighiales. Também são apresentados nos apêndices A e B deste trabalho dois artigos de divulgação científica/ensino de genética publicados durante o período do curso de doutorado.

CAPÍTULO 1

CURRENT KNOWLEDGE ABOUT CARYOCARACEAE Voigt (MALPIGHIALES): A SYNTHESIS BASED ON SCIENCE MAPPING AND SYSTEMATIC REVIEW APPROACHES[1]

Rhewter Nunes[2]; Natácia Evangelista de Lima[2]; Ivone de Bem Oliveira[3]; Mariana Pires de Campos Telles [2,4]

---

[1] Capítulo elaborado conforme as normas do periódico científico *The Botanical Review*.

[2] Genetics & Biodiversity Laboratory (LGBio), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, GO, Brasil
[3] Blueberry Breeding and Genomics Lab, Horticultural Sciences Department, University of Florida, Geinesville, FL, USA
[4] School of Agrarian and Biological Sciences, Pontifical Catholic University (PUC Goiás), Goiânia, Brazil.

Title:

Current knowledge about Caryocaraceae Voigt (Malpighiales): a synthesis based on science mapping and systematic review approaches

Authors:

Rhewter Nunes[1,*], Natácia Evangelista de Lima[1], Ivone de Bem Oliveira[2], Mariana Pires de Campos Telles[1,3]

Affiliations:

[1] Genetics & Biodiversity Laboratory (LGBio), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, GO, Brazil;

[2] Blueberry Breeding and Genomics Lab, Horticultural Sciences Department, University of Florida, Geinesville, FL, USA;

[3] School of Agrarian and Biological Sciences, Pontifical Catholic University (PUC Goiás), Goiânia, GO, Brazil.

[*] Correspondence: rhewter@gmail.com.

Keywords: *Anthodiscus*, *Caryocar*, Neotropical Biodiversity, Scientometrics.

ABSTRACT

Caryocaraceae is a family of plants widely distributed throughout the Neotropic region. It is divided into two botanical genera, *Anthodiscus* and *Caryocar*, and has a total of 26 currently accepted species. *Caryocar* is the typical genus and usually presents species with fruits much appreciated by human populations living in the regions where the natural populations of these species occur. Some of these species are of great cultural importance to traditional and indigenous populations and studies suggest that these peoples have generated a process of domestication that can be considered as initial for some species in the *Caryocar* genus. Many studies have been conducted, mainly by Brazilian researchers, in order to characterize the physical, chemical and biochemical aspects of these plants and the search for incorporation of some of these species in traditional cultivation and breeding systems. Proprieties like antibiotic, anticonvulsant, anti-inflammatory, antinoceptive, antioxidant, healing, hypolipenant as well as benefits in cardiovascular control, gastrointestinal protection, and prevention of some cancers have been reported showing the importance of this species like genetic resources.

1. INTRODUCTION

Caryocaraceae Voigt (1845) is a botanical family belonging to the order Malpighiales. This family is made up of two genus: *Anthodiscus* G.Mey. and *Caryocar* L. The genus *Anthodiscus* has ten species currently accepted: *A. amazonicus* Gleason & A.C.Sm., *A. chocoensis* Prance, *A. fragrans* Sleumer, *A. klugii* Standl. ex Prance, *A. mazarunensis* Gilly, *A. montanus* Gleason, *A. obovatus* Benth. ex Wittm., *A. peruanus* Baill., *A. pilosus* Ducke, and *A. trifoliatus* G.Mey. The genus *Caryocar* has 16 species currently accepted: *C. amygdaliferum* Mutis, *C. amygdaliforme* G.Don, *C. brasiliense* A.St.-Hil., *C. coriaceum* Wittm., *C. costaricense* Donn.Sm., *C. cuneatum* Wittm., *C. dentatum* Gleason, *C. edule* Casar., *C. glabrum* (Aubl.) Pers., *C. gracile* Wittm., *C. harlingii* Prance & Encarn., *C. microcarpum* Ducke, *C. montanum* Prance, *C. nuciferum* L., *C. pallidum* A.C.Sm., and *C. villosum* (Aubl.) Pers (Kew Botanic Gardens, 2019).

In taxonomic aspects, Caryocaraceae does not present a well clade formation with other species of the order Malghiales(APG IV et al., 2016). The most recent study in this regard suggests a polytomy formed between the clades of Chrysobalanoids, Malpighioids, Putranjivoids and Caryocaraceae since no high support values were obtained for the nodes to solve this polytomy. Thus, the exact position of Caryocaraceae within the Malpighiales order remains uncertain (Wurdack & Davis, 2009; Xi et al., 2012).

Considering the IUCN Red List, three species of *Anthodiscus* and eight species of *Caryocar* appears in the list. For *Anthodiscus* genus, *A. montanus* appears classified as “Endangered”, *A. chocoensis* as “Vulnerable” and *A. amazonicus* as “Least Concern”. For Caryocar genus we have three species in the “Endangered” category: *C. costaricense*, *C. coriaceum* and *C. amygdaliforme*. Furthermore, *C. brasiliense*, *C. villosum*, *C. glabrum* and *C. gracile* appears as “Least Concern” status. *C. nuciferum* have “Data Deficient” status (The International Union for Conservation of Nature’s Red List of Threatened Species, 2019).

The species belonging to the Caryocaraceae family have a wide distribution among the Neotropics, all of which are native to this biogeographic region. Some species of the genus *Caryocar* have edible fruits, with good quality wood and herbal effects. They are often widely known by local human populations and their fruits are popularly known as: “pequi”, “piqui”, “pequiá”, “pekea”, “pequi-vinagreiro”, “pequia-rana”, “souari” or “sawarri”. One of the species of the genus *Caryocar*, *C. brasiliense*, was listed in a list of species for the future made by the Brazilian Ministry of Environment. Potential use refers to fruits as food and oil extraction, roots

and leaves as medicinal, and trunk as wood. However, most of the fruits used for human consumption come from extractive practices, which can damage the persistence of natural populations of these species (Prance & Silva, 1973).

Only one work sought to synthesize information about the Caryocaraceae family so far (Prance, 2014). It was published in the book "Flowering Plants. Eudicots: Malpighiales" in the form of a short review chapter and serves as an important source of general knowledge about the general biology, especially botanical aspects, of the Caryocaraceae family. Although is an important reference for an overview of Caryocaraceae, some aspects are not addressed by Prance (2014), such as genetic studies with *Caryocar* species and studies related to traditional knowledge. In this sense, here we present a systematic review aiming to trace the current state of scientific knowledge about Caryocaraceae contained in the Web of Science and Scopus databases. Additionally, we performed a science mapping analysis to trace bibliographic indicators of scientific knowledge production about species belonging to the Caryocaraceae family.

2. MATERIAL AND METHODS

2.1. Data survey and selection

To retrieved the publications about Caryocaraceae family we performed a systematic review following the PRISMA statement for reporting systematic reviews and meta-analyses. We performed surveys in Scopus (https://www.scopus.com/home.uri) and Web of Science (http://apps.webofknowledge.com) databases using the combined keywords: “Caryocaraceae”, “*Caryocar*” and “*Anthodiscus*”. We exclude review and commentary papers as well as early versions of papers that have gone through the correction of their publications. The search was carried out considering all the publication period of the database until July 2019 (1988 for Scopus and 1945 for WoS). An inspection on the title, abstract and keywords of each paper was performed and papers that did not belong to the scope of this study were removed. To know about the distribution of Caryocaraceae species we searched the GBIF database and plotted the occurrences against a level II ecoregion shape.

2.2. Scientometric analysis

We quantify the scientific knowledge produced on the Caryocaraceae family using bibliometrix R package. Only the papers present in Scopus were considered for the scientometric analysis, taking into account that: i) most of the works are present in both databases; ii) the

attribution of the number of citations is heterogeneous between the databases and iii) Scopus (usually) was able to resort more articles than WoS. An additional visual inspection was performed to see if any duplicates were not detected by the function of the bibliometrix package. The descriptive metrics of the dataset were performed considering the 10 main results of each of them. In the collaborative network analysis between authors, we used the Waltrap clustering algorithm normalized by the Jaccard distance. The 50 authors with the highest values of betweenness centrality and at least one edge among authors were chosen to participate in the network.

2.3. Paper classification

In order to facilitate the collection of information in the articles related to the family Caryocaraceae, the papers obtained were classified into the following analysis groups: Anatomy and morphology; Distribution and habitat; Ecology; Genetics; Physical, chemical and biochemical characterization; Farming and industry; and Culture and traditional knowledge. Papers could be classified in more than one group, when relevant.

3. RESULTS AND DISCUSSION

3.1. Scientific knowledge production and organization

The search in Scopus database retrieved a total of 404 documents that after passing the filtering criteria resulted in a final set of 386 papers. The first article obtained in the search was published in 1938. The papers were published in a total of 208 scientific journals, with an average of 15.44 citations per document. Most of the papers were published from 2006 and the production peaks were in the years 2011 and 2013. The production of scientific knowledge related to Caryocaraceae can still be considered fickle and an annual publication constancy cannot be observed. The number of papers per author is relatively low (0.272 papers per author from a total of 1421 authors) indicating that most researchers do not have many published studies with the Caryocaraceae group and can be considered sporadic authors.

An overview of the organization of scientific knowledge related to the Caryocaraceae family can be viewed in the three-fields plot below (Figure 1). Although it is only a descriptive and visual analysis it is possible to observe which the main authors of the theme, which keywords were most used by them and in which scientific journals their works were published. *Caryocar brasiliense* and *C. coreaceum* are the only species in the Caryocaraceae family to appear in the figure. The most common keyword is "pequi", the popular name commonly attributed to the fruits

of the species of the genus Caryocar. It is also possible to observe that most publications are concentrated in Brazilian journals and deal mainly with botany and agriculture. Considering the keywords that appear in the three-fields plot it is also possible to infer that most of the work focuses on Cerrado species while the Amazonian species (which includes the entire genus *Anthodiscus*) are more neglected.

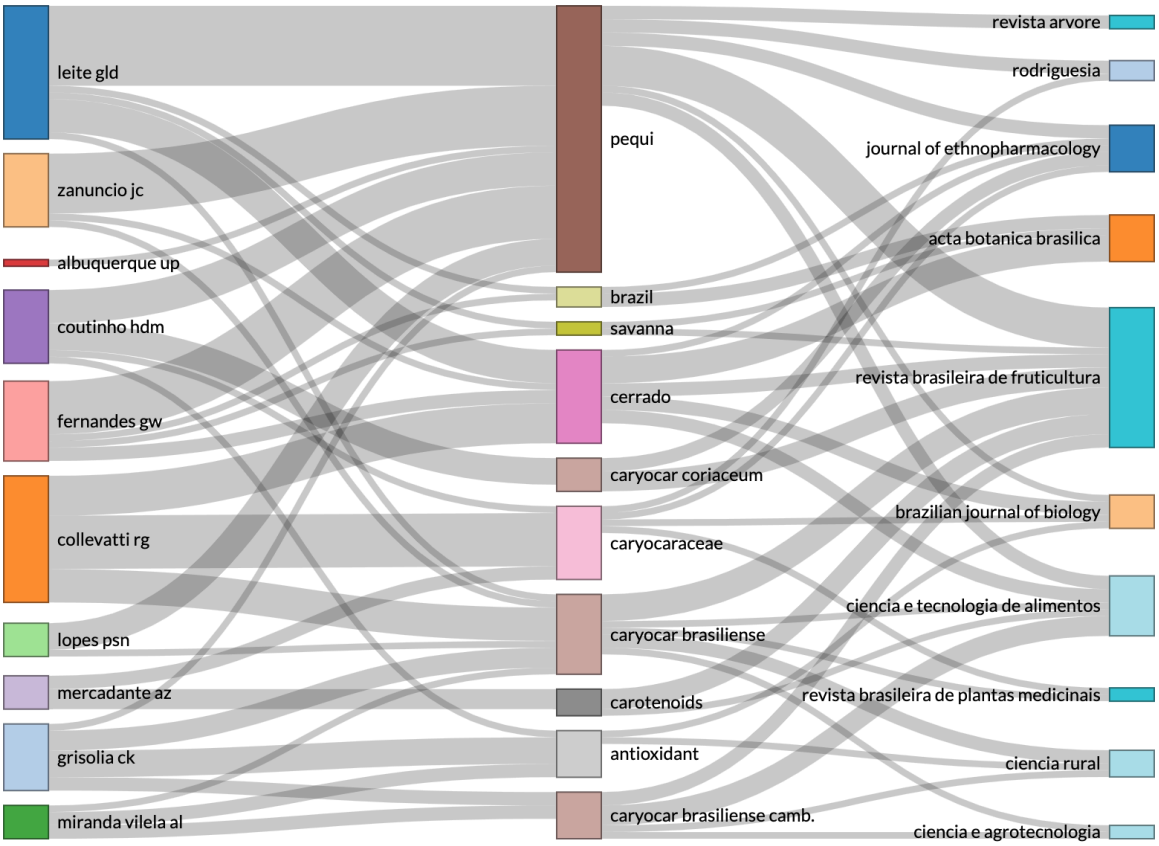


Figure 1. Three-fields plot of the production of scientific knowledge about the Caryocaraceae family available on Scopus. The rectangles represent the main authors, keywords and sources (scientific journals), respectively.

The researchers who published the most scientific papers related to the Caryocaraceae family were Leite GLD (n= 18; 4.66 %), Grisolia CK (n= 14; 3.62 %) and Collevatti RG (n= 13; 3.36 %), respectively (Figura 2a). These authors come from the following institutions, respectively: Federal University of Minas Gerais (UFMG), University of Brasilia (UnB) and Federal University of Goiás (UFG). Regarding the relevance of the authors (here measured by the authors' h-index

within the subset of papers related to the Caryocaraceae family) the three most prominent are Collevatti RG (h-index = 10), Grisolia CK (h-index = 9) and Miranda Vilela AL (h-index = 9) (Figure 2b). The three most relevant authors come from the following institutions: UFG, UnB and UnB, respectively. These results indicate the important role of Brazilian public universities in the development of work with the Caryocaraceae family. All the institutions mentioned above are federal public universities and have their campuses where the predominant biome is the Cerrado, where *Caryocar brasiliense* is very abundant and popular.

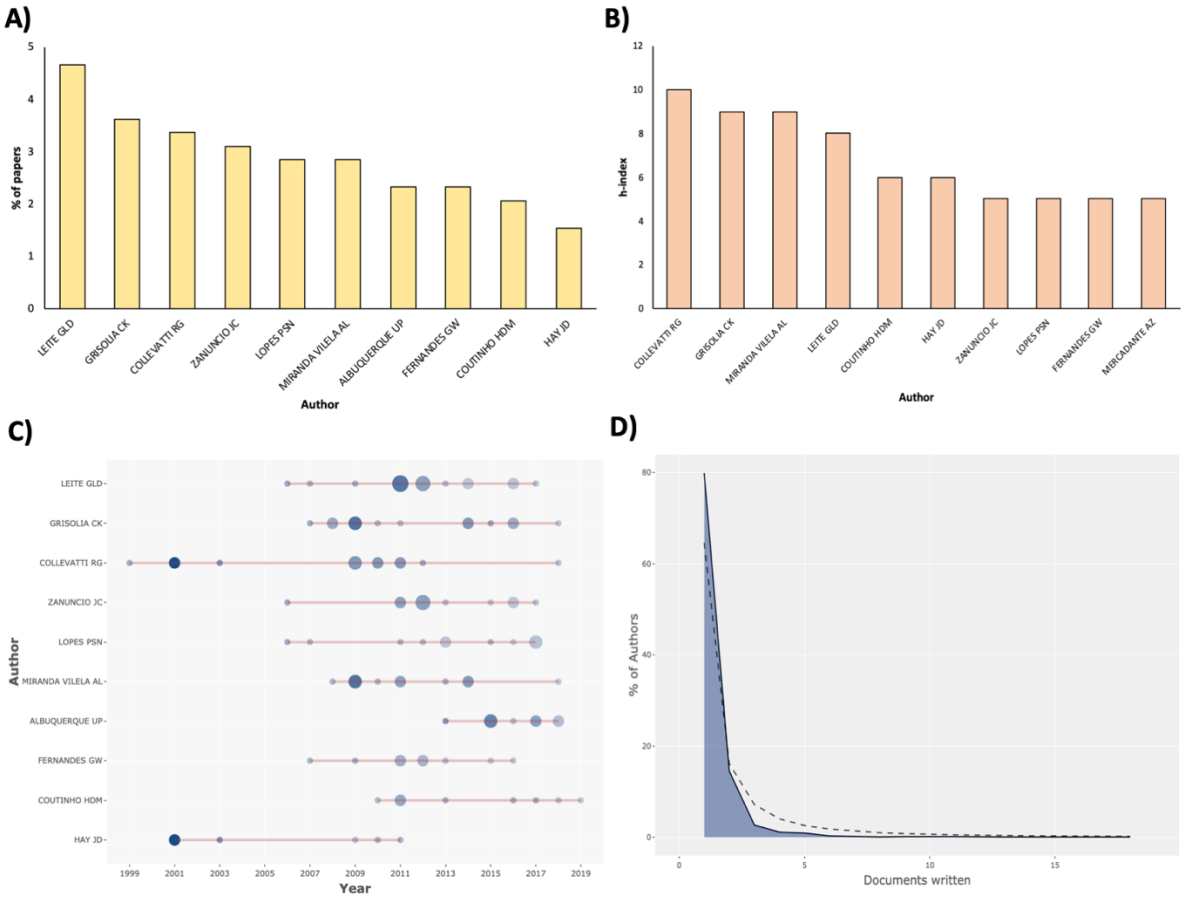


Figure 2. Authors overview of the production of scientific knowledge about the Caryocaraceae family available on Scopus. A) Distribution of paper proportion by top 10 most productive authors; B) Distribution of the top 10 authors with the highest impact scientific production; C) Top 10 authors productivity over the time. Dot size represents number of publications; and D) Author productivity through Lotka's Law.

The top 10 most relevant papers considering number of citations is showed in Table 1. Papers were published between 1984 and 2008. The total number of citations in the top 10 papers ranged from 106 for the work of Gottsberger et al. (1984) to 254 for the work of Cuevas et al. (1988). Guimarães et al. (2008) is the third work on the list and has the highest number of citations per year. The most relevant articles on scientific knowledge related to the Caryocaraceae family can be classified into two main groups on the subject: "Ecology and/or Evolution" and "Food Science".

Table 1. The top 10 most cited papers with scientific knowledge related to the Caryocaraceae family available on Scopus database. * TC = Total citations.

First author	DOI	Journal	Year	TC*	TC per Year
Cuevas E	10.1007/BF00379956	Oecologia	1988	254	8,19
Bucci SJ	10.1046/j.0140-7791.2003.01082.x	Plant, Cell & Environment	2003	217	13,56
Guimarães JR	10.1371/journal.pone.0001745	Plos One	2008	182	16,55
Alencar JW	10.1021/jf00120a031	Journal of Agricultural and Food Chemistry	1983	145	4,03
Oliveira OS	10.1046/j.1365-2435.1997.00087.x	Functional Ecology	1997	144	6,55
Roesler R	10.1590/S0101-20612007000100010	Ciência e Tecnologia de Alimentos	2007	128	10,67
Abreu FR	10.1016/j.molcata.2003.08.003	Journal of Molecular Catalysis A: Chemical	2004	124	8,27
Azevedo-Meleiro CH	10.1016/j.jfca.2004.02.004	Journal of Food Composition and Analysis	2004	117	7,80
Collevatti RG	10.1046/j.1365-294X.2001.01226.x	Molecular Ecology	2001	111	6,17
Gottsberger G	10.1007/BF00984031	Plant Systematics and Evolution	1984	106	3,03

Most of the ten most productive authors have been publishing their work on Caryocaraceae for the past ten years (Figure 2c). Collevatti RG and Hay JD appear as the pioneer researchers on the subject considering the most productive authors on a scale of the last few years. Collevatti RG is the author with the greatest time span of publication between 1999 and 2019. Productivity of authors follows Lotka's Law (most scientific knowledge of an area is produced by a few authors). Approximately 80% of the authors produced only one paper and 94.4% produced up to two papers from the total set (Figure 2d). This result reinforces the hypothesis that most researchers who published works related to the Caryocaraceae family are "sporadic researchers" in this research theme and have no tradition in publishing their work with species of this group.

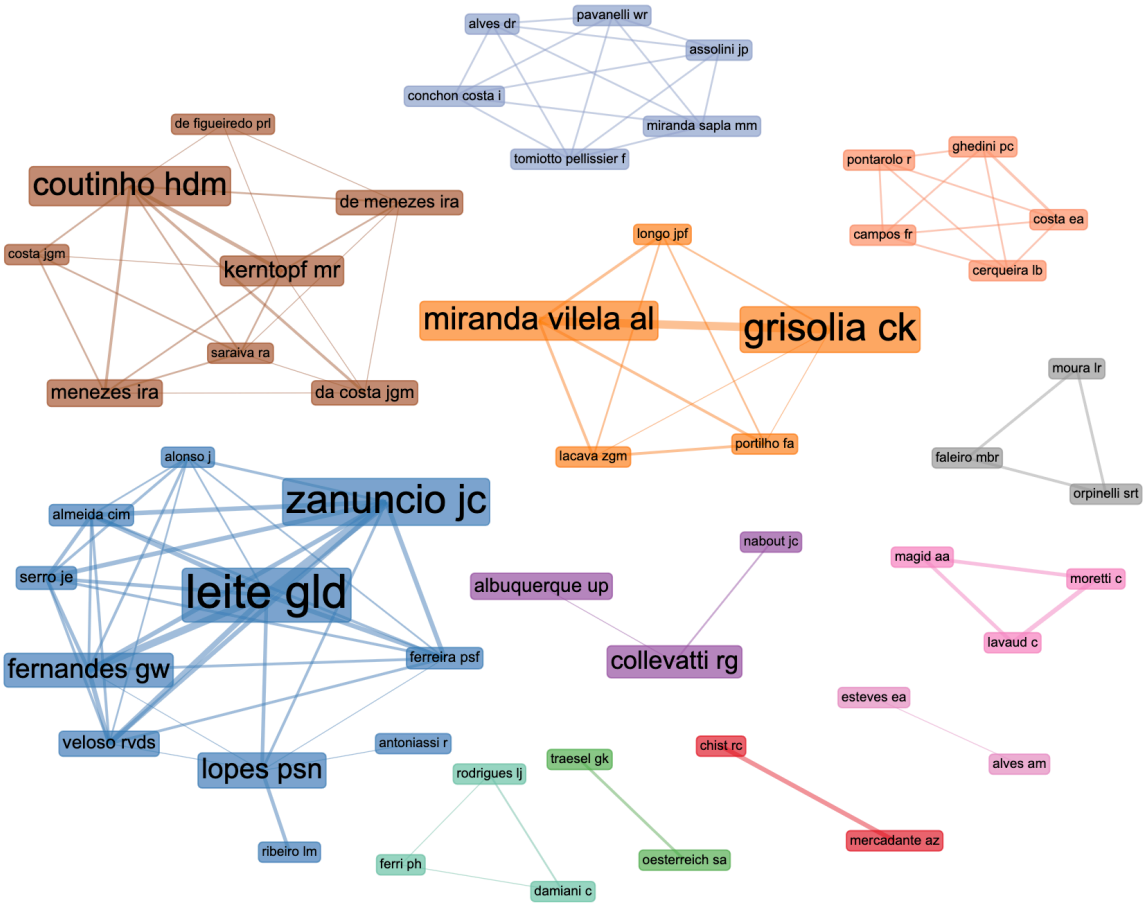


Figure 3. Collaboration network among authors who have published scientific papers related to the species of the Caryocaraceae family.

Collaboration network among authors presented the formation of 12 collaboration group clusters (Figure 3). Many of these groups had a star-shaped network structure demonstrating the co-occurrence of authorship among most or all of the authors that make up the group. For many of these groups, the authors who appeared with the most productive and/or most relevant within the Caryocaraceae-related papers group appear with the highest values of betweenness centrality (Example: GLD Milk in the dark blue cluster, Grisolia CK and Miranda Vilela AL in the orange cluster, and Collevatti RG in the purple cluster). The higher values of centrality presented by these authors can serve as a measure of leadership within the obtained clusters.

The paper classification result in: Anatomy and morphology (n= 19); Distribution and habitat (n= 83); Ecology (n= 81); Genetics (n= 21); Physical, chemical and biochemical characterization (n= 224); Farming and industry (n= 85); and Culture and traditional knowledge (n= 17). Only eight papers deal with the genus *Anthodiscus* and they relate to floristic surveys and anatomical descriptions.

3.2. Anatomy and morphology

Caryocaraceae is formed mainly by trees and may also occur in the form of shrubs more rarely (APG IV et al., 2016). Pollen: The morphology of pollen grains varies considerably by genus. *Anthosdiscus* presents pollen grains ranging from 25 to 44 μm (polar diameter) and 12 to 32 μm in width (equatorial diameter). The shape may be prolate or occasionally subprolate and the grain surface is reticulate. Usually the grains are tricolporate or rarely tetracolporate. May have some abnormally large grains (Barth, 2015). *Caryocar* presents pollen grains ranging from 36 to 120 μm (polar diameter) and 30 to 114 μm in width (equatorial diameter). The shape may be subprolate or prolate spheroidal and the grain surface is reticulate or ornate. Usually the grains are tricolporate or rarely dicolporate or tetracolporate. A particular feature of the *Caryocar* genus is the presence of prominence between the furrows and the equatorial region (Barth, 2015).

Flowers: The flowers are large, actinomorphic and hermaphrodite. The corolla is made up of four or five petals (most common) and may rarely have more than that. The chalice follows the same pattern generally having 5 sepals. The petals are imbricate and deciduous. In *Anthodiscus* they form a calyptra (Dickison, 1990). The flowers have 55 to 750 stamens and the filaments usually attach to a ring at the base of the flower. The fillets of the stamens are usually long but short fillets of sterile stamens may appear. The anthers are bilocular and can be basified or attached to the middle. The ovary is rough and has four to six carpels in *Caryocar* and eight to 20 in *Anthodiscus*. The eggs are basal and atropic (Dickison, 1990).

Fruits: In both genera the fruits are drupa type with one to four seeds in *Caryocar* and 8 to 20 in *Anthodiscus*. The mesocarp is indiscent and thick. The endocarp is hard and spiny. On the outside it has a kind of internal mesocarp attached to the endocarp forming the pyrenes that are individually divided by seed (Prance & Silva, 1973). Seeds: The seeds are commonly reniform, with two cotyledons and thin or absent endosperm. The embryo may have a straight, arched or spiral root and a fleshy hypocotyl (Prance & Silva, 1973).

Leaves: Brochidodromous and camptodromous nerves are the usual patterns, although some species of *Anthodiscus* are hyphodromous. The leaf mesophyll and petioles have branched sclerenchymal idioblasts. Most stomata are mostly anomocytic but anisocytic or paracitic forms can also occur. They are more concentrated in the abaxial part of the leaves (Ramos et al., 2015). Wood: *Anthodiscus* and *Caryocar* present many differences regarding the morphological characteristics of the wood. Vessel abundance is quite different between genders, being higher in *Anthodiscus* (15 mm^2) than in *Caryocar* (3 mm^2). *Anthodiscus* vessels have an average diameter of 50 to 100 mm and can be solitary or in multiples of two to six cells. In *Caryocar* they range from 74 to 577 mm in diameter and are solitary or in multiples of two to six cells. The specific gravity of the wood ranges from 0.802 to 0.906 and can be considered from moderately hard and heavy to extremely hard and heavy. It can be considered durable and moderately easy to work on most species (Prance & Silva, 1973).

3.3. Distribution and habitat

The distribution of Caryocaraceae individuals based on GBIF occurrence data show that this is a neotropical family that occurs from northern Costa Rica to southern Brazil (Figure 4). The species of the genus *Anthodiscus* occur mainly in the northwestern portion of South America. However, this genus has a disjoint distribution, also occurring, to a lesser extent, in the Atlantic Forest region of Bahia (coast of Brazil) without occurring in the regions that connect Bahia with the Amazon (such as savanna ecoregion, for example). Many of these species have a predominantly restricted distribution such as *Anthodiscus peruanus* and *Anthodiscus mazarunensis*. The genus *Caryocar* has a wider occurrence within the distribution area of the Caryocaraceae family.

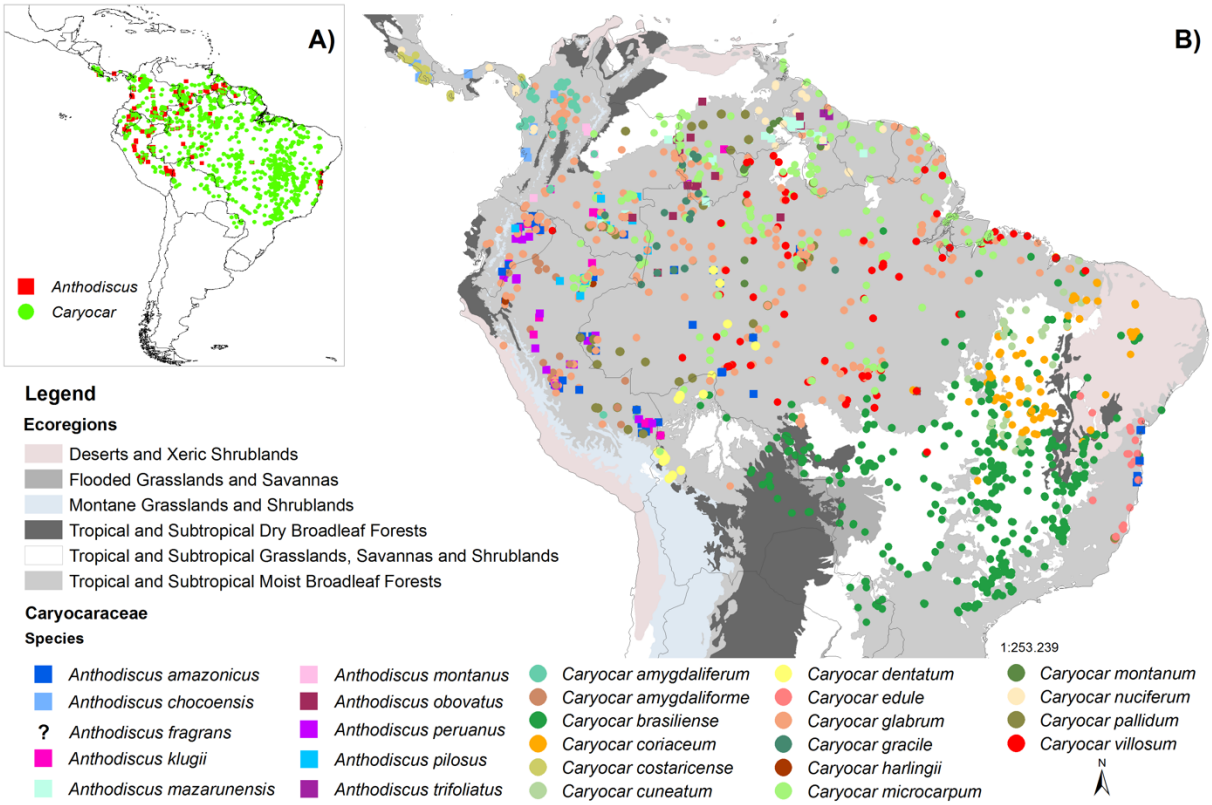


Figura 4. Distribution of Caryocaraceae based on occurrence GBIF data. A) Distribution considering the two genera of Caryocaraceae family: *Anthodiscus* and *Caryocar*; and B) Distribution considering each species of Caryocaraceae family.

3.4. Ecology

For *Caryocar brasiliense*, pollination is performed by glossophagine bats (*Glossophaga soricina* and *Anoura geoffroyi*) (Gribel & Hay, 1993). Non-glossophagine bat species *Phyllostomus discolor*, *Vampyrops lineatus* and *Carollia perspicillata* may be occasional pollinators. Sphingidae insects, such as *Erinyis ello* and *Pseudosphinx tetrio*, can also be occasional pollinators (Gribel & Hay, 1993). Based on the morphology of the flowers it is very likely that other species of the genus *Caryocar* are also pollinated by bats and evidence of this has already been observed for *C. villosum* (Prance & Silva, 1973). For *Anthodiscus* there are no studies on possible pollinators but flower morphology suggests that this is more likely to occur by insects than by bats (Prance & Silva, 1973).

Dispersion is performed by agoutis (*Dasyprocta ssp.*) or rheas (*Rhea americana*) (Prance, 2014). An important issue regarding the ecology of dispersal of species belonging to the Caryocaraceae family is that there was probably a decline in the number of natural dispersers with

megafauna extinction events. The seeds of this group are usually large and heavy requiring dispersers that support this load for gene flow to occur over long distances. Rodent dispersal, flooding, gravity and human-mediated dispersion are probably the way these species have dispersed today (Guimarães, Galetti, & Jordano, 2008).

3.5. Genetics

Studies on the genetic aspects of the Caryocaraceae family are limited to the genus *Caryocar* and are mainly related to the evaluation of the genetic diversity of natural populations. Nevertheless, basic information such as karyotype characterization was performed for *Caryocar brasiliense*, *C. microcarpum* and *C. villosum*, presenting 2n = 46 for all this species and suggests that they are paleohexploid species (Ehrendorfer, Morawetz, & Dawe, 1984). Genome size estimates such as C-values or any other chromosomic/cytogenetic information were not found for any Caryocaraceae species. Efforts to generate genetic information were made for *C. brasiliense*. A protocol for extracting genomic DNA from leaf tissue (Silva, 2010), molecular markers based on microsatellite region polymorphism (Collevatti, Brondani, & Grattapaglia, 1999), and DNA sequences for phylogeographic studies have been established for this species (Collevatti, Grattapaglia, & Hay, 2003; Collevatti, Leoi, et al., 2009). The ten microsatellite loci were developed using genomic enrichment libraries and had a number of alleles per locus ranging from ten to 22. These loci were successfully transferred to *C. coriaceum*, *C. edule*, *C. glabrum*, *C. pallidum* and *C. villosum* (Collevatti, Brondani, & Grattapaglia, 1999) and present a high combined probability of paternity exclusion (0.9999). Also, sequences for application in phylogeographic studies were generated for the chloroplast regions of *C. brasiliense* trnT, trnF, trnL and psbA-trnH (Collevatti, Grattapaglia, & Hay, 2003; Collevatti, Leoi, et al., 2009).

The characterization of genetic diversity of *Caryocar* populations has already been performed using both phenotypic and molecular markers. Genetic diversity of traits related to early growth was estimated in *C. brasiliense* populations (Moura et al., 2013; Santos et al., 2018). The two studies focused on initial growth traits were evaluated for plant height, stem diameter at ground level and crown diameter (Santos et al., 2018); and seed emergence percentage, seedling emergence time, height and diameter growth rate, final height, final diameter and emergence seedling survival rate (Moura et al., 2013). The results of these studies reveal the existence of great genetic variability for the studied characters with great population effect. The effect of population differentiation is probably related to heterogeneity in germination rates. The recommendation in both papers is that conservation of genetic resources thinking of early growth traits should be directed to many populations (Moura et al., 2013; Santos et al., 2018).

In another study, the evaluation of genetic variability of fruit characteristics of *C. brasiliense* reveals no significant effect of populations (Santos et al., 2018). The evaluated characters were total fruit mass, external mesocarp mass per fruit, number of putamens (drupes) per fruit, total mass of putamens per fruit, average mass of putamens per fruit, total pulp mass per fruit and average pulp mass per fruit. Opposite to early growth traits, the conservation of genetic resources thinking about fruit characteristics should take into account the collection of many individuals from few populations (Santos et al., 2018).

The molecular genetic diversity caracterization of *Caryocar* species was performed using diferent markers such as isoenzymes/aloenzymes (Afrânio Farias de Melo-Júnior et al., 2004), RAPD (Random Amplified Polymorphic DNA) (Londe et al., 2010) and SSR (Simple Sequence Repeats) (Collevatti, Brondani, & Grattapaglia, 1999). Ten isoenzymatic sets were used to make estimates of genetic diversity in 60 individuals from four subpopulations of *C. brasiliense* where a low genetic divergence was observed among the studied populations ($\theta = 0.020$) and absence of inbreeding ($f = -0.449$) (Melo-Júnior et al., 2004). In a later study with a larger number of individuals (240), there was a persistence of the result of low genetic divergence among populations ($\theta = 0.036$) (Melo-Júnior et al., 2012). The low values of genetic differentiation obtained in these studies are probably associated with a restricted sampling of geographically close populations, which may facilitate the genetic flow and the consequent decrease in genetic differentiation among populations.

RAPD markers were applied to study the genetic differentiation of *C. brasiliense* individuals from populations with and without thorns in the stone (Londe et al., 2010). This analysis was performed on 20 individuals from five populations, one of them containing no thorn in the stone. It was observed that the existence of genetic divergence among the population without thorn in the stone compared with those having the thorn stone. Sequencing of the banding regions and sequence analysis showed the similarity of the Dof1 genes and the *Zea mays* phosphinothricin acetyltransferase gene (a regulator for C4 photosynthetic phosphoenolpyruvate carboxylase gene expression) (Londe et al., 2010).

The set of ten microsatellite markers developed by Collevatti, Brondani & Grattapaglia (1999) were used to answer several questions about genetic diversity, reproductive biology and ecology of the *Caryocar* genus, especially for *C. brasiliense*. These loci were used to characterize the genetic diversity of 314 from 10 subpopulations of *C. brasiliense* revealing the presence of inbreeding ($f = 0.11$) and the genetic differentiation between populations correlated with increasing distance (Mantel test correlation, $r = 0.518$, $p < 0.05$) (Collevatti, Grattapaglia, & Hay, 2001).

The set of SSR markers was also used to genotype *C. brasiliense* matrices and progenies from four subpopulations in a mating system investigation (Collevatti, Grattapaglia, & Hay, 2001). The obtained results indicate that *C. brasiliense* has a mixed crossing system, presenting a multilocus cross fertilization rate of 1 and by locus ranging from 0.769 and 0.869 according to the population. Considering that the difference between single locus and multilocus cross-fertilization rates is different from zero, one can consider the possible occurrence of biparental crossing and consequent inbreeding due to the isolation of populations due to the low availability of seed dispersers and pollinators (Collevatti, Grattapaglia, & Hay, 2001).

In another study, this time with an intrapopulation approach, pollen dispersion and crossing structure in *C. brasiliense* was evaluated (Rosane G. Collevatti et al., 2010). This study reinforces that *C. brasiliense* has a mixed breeding system (multilocus cross-fertilization rate of 0.891) and a high probability of occurrence of complete siblings (same pollen and egg donor) between offspring of the same matrix (rp = 0.135). It was also possible to observe a maximum pollen dispersion distance of 500 m and 80% of pollen dispersion events occur at distances of less than 200m. In addition, a significant but weak spatial genetic structure was observed (Sp = 0.0116) (Collevatti et al., 2010).

Although the *C. brasiliense* crossing system is mixed, cross fertilization seems to be an important event in the persistence of populations of this species. The aborted seeds in *C. brasiliense* fruits come from self-fertilization or cross-fertilization in which the pollen donor is very genetically close to the matrix that originated the fruit (Collevatti, Estolano, et al., 2009). This phenomenon highlights the relationship between inbreeding depression and seed abortion in *C. brasiliense*. It is indicated that a fragmentation may endanger the natural populations of *C. brasiliense* if the fragments impede the gene flow between them. Another factor that is contributing to the decrease in gene flow is the low availability of pollinators and dispersers (Collevatti, Estolano, et al., 2009). In a comparative analysis of spatial genetic structure among three species that occur in the Cerrado (*Caryocar brasiliense*, *Dipteryx alata* and *Tibouchina papyrus*) it was observed that in species that are dispersed by mammals the spatial genetic structure is larger than that of wind dispersed species (Collevatti et al., 2010). A demographic and kinship analysis of a population of *C. brasiliense* over a 23-year period reveals that the probability of death correlates with the individual genotype. The lower the heterozygosity, the greater the likelihood of death. It was also observed that the relationship is correlated with the spatial distance between pairs of individuals less than 10m, and the relationship structure has not changed throughout the life stages (Collevatti & Hay, 2011).

A deeper analysis of the processes that generate genetic structuring of 10 *C. brasiliense* populations was performed using an integrative approach involving estimates of genetic diversity, ecological niche modeling and landscape genetics (Diniz-Filho et al., 2009). The results indicate that genetic diversity tries to obey a center-periphery model. It was also possible to observe a decrease in diversity values in southern populations of the species distribution area, which may be related to areas of greater human occupation and fragmentation (Diniz-Filho et al., 2009). Diniz-Filho et al. (2009) emphasize the need to analyze more populations for a more accurate understanding since 10 local populations may be considered little for this type of approach.

The evolutionary history of *C. brasiliense* was investigated using chloroplast region sequences and microsatellite markers in a phylogeographic approach (Collevatti, Grattapaglia, & Hay, 2003). Evidence was found that multiple strains formed the current *C. brasiliense* populations. These results provide evidence for the hypothesis of restricting populations to wet refuges in times of prolonged drought followed by dissemination throughout central Brazil during the interglacial period. In addition, it is also possible to hypothesize that the possibility of gene flow via seed has become infrequent due to the extinction of megafauna in the last glaciation, causing gene flow between populations to be maintained mainly via pollen (Collevatti, Grattapaglia, & Hay, 2003). The demographic history of *C. brasiliense* in an approach involving ecological niche modeling and coalescence analysis was also analyzed (Collevatti et al., 2012). Three chloroplast regions (psbA-trnH, trnC-ycf6 and trnL intron) were used which showed low genetic diversity and high differentiation between populations. The results show a wider current distribution than estimated for the quaternary dry periods and favor the hypothesis of formation of multiple refuges due to shrinkage in populations at the last glacial maximum (21 kyr ago) (Collevatti et al., 2012).

Another approach also using ecological niche modeling was undertaken to try to predict how the genetic diversity behavior of *C. brasiliense* populations will respond in response to predicted climate change over the next 50 years (Collevatti, Nabout, & Diniz-Filho, 2011). The results suggest that the area of climate suitability for the species will be restricted in the currently more inbreeding populations (populations further south of the species distribution) and which are more fragmented. This change in the range of suitability of the species may lead to the loss of great genetic diversity and may endanger the natural populations of *C. brasiliense* (Collevatti, Nabout, & Diniz-Filho, 2011).

Considering other species of the genus Caryocar, the analysis of the genetic structure between populations of *C. microcarpum* (species that occurs in flooded areas) and *C. villosum* (species that occurs in dry area) in the Rio Negro region to test the hypothesis that The Negro river

has served as a barrier to gene flow for the populations of these two species (Collevatti, Leoi, et al., 2009). Two chloroplast regions (intron of the trnL gene and the intergenic region psbA-trnH) and 10 SSR loci were used. For *C. microcarpum* a total of 13 haplotypes were obtained for the sequenced regions (6 for the trnL intron and 7 for psbA-trnH) while for *C. villosum* only one haplotype was obtained for the two regions under analysis. The results suggest that the Rio Negro does not act as a barrier to gene flow for populations on the left and right bank of the river. In addition, the results show that multiple maternal lineages formed the *C. microcarpum* populations of this region while in *C. villosum* there was a recent expansion of a maternal lineage that became refugees in the Guyana Shield during prolonged drought in the gracial period (Collevatti, Leoi, et al., 2009). The paper shows that different species of the *Caryocar* genus that co-occur in the same region probably had very different evolutionary histories from their populations.

3.6. Physical, chemical and biochemical characterization

The medicinal properties already reported in the scientific literature for species of the Caryocaraceae family are centered on species of the genus *Caryocar*. The papers results indicated of the following properties: antibiotic (Lacerda-Neto et al., 2018), anticonvulsant (Oliveira et al., 2017), anti-inflammatory (Torres et al., 2016), antinoceptive (Oliveira et al., 2015), antioxidant (Almeida et al., 2012), healing (Oliveira et al., 2010), hypolipenant (Figueiredo et al., 2016), as well as benefits in cardiovascular control (Oliveira et al., 2018), gastrointestinal protection (Lacerda-Neto et al., 2017), and prevention of some cancers (Suffredini et al., 2007).

Most of these papers are related to the antioxidant and antioxidant properties of different parts of the plant in aqueous extracts or pequi oil. Pequi oil (*C. brasiliense*) intake and its potential antioxidant effect on runner blood pressure was evaluated in a study by Miranda-Vilela et al. (2009). It was observed that pequi oil functioned as an anti-inflammatory agent against the effects of intense physical activity and also functioned as a modulator of total cholesterol, especially in men. It has also been observed that pequi oil decreases blood pressure regardless of gender (Miranda-Vilela et al., 2009). In another paper, runner athletes were also used as populations in the study of interleukin gene polymorphism (IL-6) and pequi oil supplementation (*C. brasiliense*). A significant difference between GC and CC (mutant) genotypes was observed in exercise-induced damage and C-reactive protein (CRP) levels. These results reveal that although there is evidence of the benefits of pequi oil supplementation in high performance athletes, there may be a fluctuation in the beneficial effect related to the athlete's genetic background (Miranda-Vilela, Ribeiro, & Grisolia, 2016).

A study aimed to evaluate the effect of pequi (*C. brasiliense*) oil in conjunction with dextran-functionalized magnetic fluid (DexMF) in the treatment of advanced clinical Ehrlich-solid-tumor (Miranda-Vilela et al., 2013). It has been observed that pequi oil treatment is efficient in the tumor cell necrosis process, especially after the second week of treatment. In another study it was observed that aqueous extract of *C. microcarpum* acts against human central nervous system cancer tumor cells (cell line SF-268) (Suffredini et al., 2007). Both studies indicate the potential use of *Caryocar* species in cancer prevention and treatment.

The evaluation of pequi (*C. coriaceum*) almond pulp and oil indicates that they are rich in oleic and palmitic fatty acids. Because of these characteristics, Pereira et el. (2019) point out that pequi oil contains important substances to fight multi-resistant bacteria and can be used individually or as a complement to other types of antibiotics. Another biochemical study identified tannins, phenols and flavonoids in the hydroalcoholic extract produced from *C. coriaceum* leaves. A more comprehensive review of the phytochemical composition of Caryocaraceae was made by Ascari, Takahashi, & Boaventura (2013). Caryocaraceae plant compounds have also been used as a physical stabilizer in cosmetic emulsions (Raiser et al., 2018).

Pequi oil has been reported as a good microflotting agent. It can be used for apatite flotation with a 95% uptake capacity even at very low concentrations such as 2.5 mg L^{-1} (Silva, Silva, & Silva, 2015). Pequi oil has also been investigated for its ability to minimize the side effects of some types of chemotherapy (Miranda-Vilela et al., 2014).

3.7. Farming and industry

Some *Caryocar* species such as *C. brasiliense* and *C. coriaceum* are much appreciated by local human populations and are candidates for rapid incorporation into traditional cultivation systems (Grzebieluckas et al., 2010). Because of this, some studies on agronomic aspects seed germination (Sousa et al., 2017), seedling production (Santos et al., 2006), evaluation of plant morphological characters and phenology(Leite et al., 2006), and response to cultivation conditions have been performed (Françoso et al., 2014). Another growing research sector is the field of food science and technology with sensory assessments of pequi fruits, oil and other byproducts (Gonçalves et al., 2011; Vilas Boas et al., 2012; Monteiro et al., 2014; Sousa et al., 2016).

3.8. Culture and traditional knowledge

The possible occurrence of onset of *C. coriaceum* domestication was evaluated in the Chapada do Araripe region (Ceará, Brazil) (Sousa Júnior et al., 2018). It was observed significant difference in fruit size of *C. coriaceum* under cultivation when compared to trees that were only

had fruits collected from natural populations (Sousa Júnior et al., 2018). Indication of *Caryocar sp.* domestication was also observed by archaeological evidence from southwestern Amazonia (Watling et al., 2018). Archaeological data indicate landscape transformation by planting in this region dating from the Middle Holocene (Watling et al., 2018). Such studies provide evidence for a recent domestication process for some *Caryocar* species in different regions of South America.

Some studies have been conducted on the traditional knowledge associated with some species of the genus *Caryocar* (Sousa-Júnior, Albuquerque, & Peroni, 2013; Pinto et al., 2016; Smith & Fausto, 2016). Smith & Fausto (2016) investigated the different sociocultural aspects related to the cultivation and use of *Caryocar brasiliense* by the Kuikuro people in the upper Xingu River region (Mato Grosso, Brazil). Study data were collected using questionnaires and observations that were documented in photos and videos. This local is a federal indigenous reserve region and home to nine indigenous peoples of different ethnicities and languages. The obtained results indicate that the studied indigenous people practices the cultivation of *C. brasiliense* seeds in the form of orchards and that the fruit selection occurs in order to preserve the interspecific variability (based on morphology) of this plant. Planting in orchards allows the preservation of the natural populations of *C. brasiliense*, since the fruits are not the target of extractivism. It is also observed that pequi is such an important fruit for this human population that it can be considered a "biocultural" plant in the sense that it has importance as a genetic resource as in the social aspect (Smith & Fausto, 2016).

Traditional knowledge regarding the use of *C. brasiliense* was also characterized in a “quilombola community” (descendants of these enslaved peoples during the slavery period) of Pontinha (Minas Gerais, Brazil) (Pinto et al., 2016). Use in food, soap production and oil production were the most cited by the population studied. In addition, the interviewees reported noticing the visit of bees in the flowers and the presence of caterpillars in the leaves. It was also commented that the fire negatively affects the natural seedlings of *C. brasiliense* leading to death (Pinto et al., 2016).

Aspects of traditional knowledge related to *Caryocar coriaceum* were also studied in the northeast region of Brazil (Chapada do Araripe, Ceará, Brazil) (Sousa-Júnior, Albuquerque, & Peroni, 2013). *C. coriaceum* fruits are very important for local human populations both for subsistence and as a source of income. All families interviewed said they market both *C. coriaceum* fruits and oil. Oil, in turn, has a higher added value and usually increases the income of these families. Knowledge about C. *coriaceum* was homogeneous among the studied families. However, there was a significant difference between knowledge about use between men and women, indicating that gender roles structure part of the knowledge about this species in the studied population. The fruit gathering process involves all family members, and even camps can be found

near *C. coriaceum* populations to accommodate the families. It is also common to hold a festival to celebrate the annual harvest, which reinforces the cultural importance of this species (Sousa-Júnior, Albuquerque, & Peroni, 2013).

The different species of *Caryocar* form a set of genetic resources that are very important for local human populations. Several studies have pointed out the importance of these species for the economy of these populations who sell or use their fruits, oil, coal or medicinal property (Azevedo, Martins, & Drummond, 2009; Santos & Mitja, 2011; Almeida et al., 2012; Ribeiro et al., 2014; Cavalcanti et al., 2015; R. R. V. Silva, Gomes, & Albuquerque, 2015; Conceição et al., 2017; Maciel et al., 2018). Regarding the artisanal extraction of pequi oil, Cavalcanti et al. (2015) produced an important photographic survey of how this process is performed.

4. CONCLUSION

- Scientific knowledge related to the Caryocaraceae family is produced by few continuous researchers and many seasonal researchers, usually Brazilian and linked to public universities.
- The most studied topics about the Caryocaraceae family are related to the physical, chemical and biochemical characterization of these plants.
- *Caryocar*, especially *Caryocar brasiliense*, are the main study targets due to its wide occurrence and economic and cultural importance.
- *Anthodiscus* is extremely lacking in scientific studies and the few existing ones are related to morphological characterization and distribution.
- Many species in the Caryocaraceae family are actual or potential genetic resources and a broader understanding of the biology of the different species that make up this group is required.

5. REFERENCES

Almeida, L.S., J.R.V. Gama, F.A. Oliveira, J.O.P. Carvalho, D.C.M. Gonçalves, & G.C. Araújo. 2012. Fitossociologia e uso múltiplo de espécies arbóreas em floresta manejada , Comunidade Santo Antônio , município de Santarém , Estado do Pará Phytosociology and multiple use of forest species in a logged forest in. Acta Amazonica 42: 185–194.

Almeida, M.R., A.F. Aissa, J.D.C. Darim, T.D.U.. Gomes, R.C. Chisté, A.Z. Mercadante, L.M.G. Antunes, & M.L.P. Bianchi. 2012. Effect of Piquiá (Caryocar Villosum) Pulp Fruit on

Oxidative Stress, Ephx2 and Tp53 Gene Expressions in Liver of Rats. Free Radical Biology and Medicine 53: S82. Available online: http://dx.doi.org/10.1016/j.freeradbiomed.2012.10.330

APG IV, (Angiosperm Phylogeny Group), M.W. Chase, M.J.M. Christenhusz, M.F. Fay, J.W. Byng, W.S. Judd, D.E. Soltis, D.J. Mabberley, A.N. Sennikov, P.S. Soltis, P.F. Stevens, B. Briggs, S. Brockington, A. Chautems, J.C. Clark, J. Conran, E. Haston, M. Möller, M. Moore, R. Olmstead, M. Perret, L. Skog, J. Smith, D. Tank, M. Vorontsova, & A. Weber. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181: 1–20. Available online: https://academic.oup.com/botlinnean/article-lookup/doi/10.1111/boj.12385

Ascari, J., J.A. Takahashi, & M.A.D. Boaventura. 2013. The phytochemistry and biological aspects of Caryocaraceae family. Revista Brasileira de Plantas Medicinais 15: 293–308. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-05722013000200019&lng=en&tlng=en

Azevedo, A.I., H.T. Martins, & J.A.L. Drummond. 2009. A dinâmica institucional de uso comunitário dos produtos nativos do cerrado no município de Japonvar (Minas Gerais). Sociedade e Estado 24: 193–228.

Barth, O.M. 2015. Estudos morfológicos dos pólens em Caryocaraceae. Rodriguesia 36: 65–71.

Cavalcanti, M.C.B.T., L.Z. de O. Campos, R. da S. Sousa, & U.P. Albuquerque. 2015. Pequi (Caryocar coriaceum Wittm., Caryocaraceae) Oil Production: A strong economically influenced tradition in the Araripe region, northeastern Brazil. Ethnobotany Research and Applications 14: 437–452. Available online: http://journals.sfu.ca/era/index.php/era/article/view/1161

Collevatti, R., R. Brondani, & D. Grattapaglia. 1999. Development and characterization of microsatellite markers for genetic analysis of a Brazilian endangered tree species Caryocar brasiliense. Heredity 83: 748–756. Available online: http://www.nature.com/doifinder/10.1046/j.1365-2540.1999.00638.x

Collevatti, R., D. Grattapaglia, & J.D. Hay. 2001. High resolution microsatellite based analysis of the mating system allows the detection of significant biparental inbreeding in Caryocar brasiliense, an endangered tropical tree species. Heredity 86: 60–67. Available online: http://www.nature.com/doifinder/10.1046/j.1365-2540.2001.00801.x

Collevatti, Rosane G., R. Estolano, S.F. Garcia, & J.D. Hay. 2009. Seed abortion in the bat

pollinated Neotropical tree species, Caryocar brasiliense (Caryocaraceae). Botany 87: 1110–1115. Available online: http://www.nrcresearchpress.com/doi/10.1139/B09-054

Collevatti, Rosane G., R. Estolano, S.F. Garcia, & J.D. Hay. 2010. Short-distance pollen dispersal and high self-pollination in a bat-pollinated neotropical tree. Tree Genetics & Genomes 6: 555–564. Available online: http://link.springer.com/10.1007/s11295-010-0271-4

Collevatti, Rosane G., D. Grattapaglia, & J.D. Hay. 2001. Population genetic structure of the endangered tropical tree species Caryocar brasiliense, based on variability at microsatellite loci. Molecular Ecology 10: 349–356.

Collevatti, Rosane G., D. Grattapaglia, & J.D. Hay. 2003. Evidences for multiple maternal lineages of Caryocar brasiliense populations in the Brazilian Cerrado based on the analysis of chloroplast DNA sequences and microsatellite haplotype variation. Molecular Ecology 12: 105–115. Available online: http://doi.wiley.com/10.1046/j.1365-294X.2003.01701.x

Collevatti, Rosane G., & J.D. Hay. 2011. Kin structure and genotype-dependent mortality: a study using the neotropical tree Caryocar brasiliense. Journal of Ecology 99: 757–763. Available online: http://doi.wiley.com/10.1111/j.1365-2745.2011.01796.x

Collevatti, Rosane G., L.C.T. Leoi, S.A. Leite, & R. Gribel. 2009. Contrasting patterns of genetic structure in Caryocar (Caryocaraceae) congeners from flooded and upland Amazonian forests. Biological Journal of the Linnean Society 98: 278–290. Available online: https://academic.oup.com/biolinnean/article-lookup/doi/10.1111/j.1095-8312.2009.01287.x

Collevatti, Rosane G., J.C. Nabout, & J.A.F. Diniz-Filho. 2011. Range shift and loss of genetic diversity under climate change in Caryocar brasiliense, a Neotropical tree species. Tree Genetics and Genomes 7: 1237–1247.

Collevatti, Rosane Garcia, M.S. Lima-Ribeiro, A.C. Souza-Neto, A.A. Franco, G. de Oliveira, & L.C. Terribile. 2012. Recovering the demographical history of a Brazilian cerrado tree species Caryocar brasiliense: Coupling ecological niche modeling and coalescent analyses. Natureza a Conservacao 10: 169–176.

Collevatti, Rosane Garcia, J.S. Lima, T.N. Soares, & M.P. de C. Telles. 2010. Spatial Genetic Structure and Life History Traits in Cerrado Tree Species: Inferences for Conservation. Natureza & Conservação 08: 54–59. Available online: http://doi.editoracubo.com.br/10.4322/natcon.00801008

Conceição, S.P., J.R.V. Gama, R.N. Monteiro, R.J.S. Ferreira, & P.S. Sousa. 2017. Production

chain in piquiá Santarém municipality , Pará State , Brazil Cadeia produtiva do piquiá no município de Santarém , Estado do Pará , Brasil. Nativa 5: 31–36.

de Figueiredo, P.R.L., I.B. Oliveira, J.B.S. Neto, J.A. de Oliveira, L.B. Ribeiro, G.S. de Barros Viana, T.M. Rocha, L.K.A.M. Leal, M.R. Kerntopf, C.F.B. Felipe, H.D.M. Coutinho, & I.R. de Alencar Menezes. 2016. Caryocar coriaceum Wittm. (Pequi) fixed oil presents hypolipemic and anti-inflammatory effects in vivo and in vitro. Journal of Ethnopharmacology 191: 87–94. Available online: http://dx.doi.org/10.1016/j.jep.2016.06.038

de Lacerda Neto, Luís J., A.G.B. Ramos, M.R. Kerntopf, H.D.M. Coutinho, L.J. Quintans-Junior, J.R.G.S. Almeida, J. Ribeiro-Filho, & I.R.A. Menezes. 2018. Modulation of antibiotic activity by the hydroalcoholic extract from leaves of Caryocar coriaceum WITTM. Natural Product Research 32: 477–480. Available online: http://dx.doi.org/10.1080/14786419.2017.1312396

de Lacerda Neto, Luis Jardelino, A.G.B. Ramos, V. Santos Sales, S.D.G. de Souza, A.T.L. dos Santos, L.R. de Oliveira, M.R. Kerntopf, T.R. de Albuquerque, H.D.M. Coutinho, L.J. Quintans-Júnior, A.G. Wanderley, & I.R.A. de Menezes. 2017. Gastroprotective and ulcer healing effects of hydroethanolic extract of leaves of Caryocar coriaceum: Mechanisms involved in the gastroprotective activity. Chemico-Biological Interactions 261: 56–62. Available online: http://dx.doi.org/10.1016/j.cbi.2016.11.020

de Oliveira, F.F.B., J.C.B. de Araújo, A.F. Pereira, G.A.C. Brito, D.V. Gondim, R.D.A. Ribeiro, I.R.A. de Menezes, & M.L. Vale. 2015. Antinociceptive and anti-inflammatory effects of Caryocar coriaceum Wittm fruit pulp fixed ethyl acetate extract on zymosan-induced arthritis in rats. Journal of Ethnopharmacology 174: 452–463. Available online: http://dx.doi.org/10.1016/j.jep.2015.08.017

de Oliveira, M.L.M., D.C.S. Nunes-Pinheiro, A.R. Tomé, É.F. Mota, I.A. Lima-Verde, F.G. de M. Pinheiro, C.C. Campello, & S.M. de Morais. 2010. In vivo topical anti-inflammatory and wound healing activities of the fixed oil of Caryocar coriaceum Wittm. seeds. Journal of Ethnopharmacology 129: 214–219. Available online: https://linkinghub.elsevier.com/retrieve/pii/S0378874110001765

Dickison, W.C. 1990. A Study of the Floral Morphology and Anatomy of the Caryocaraceae. Bulletin of the Torrey Botanical Club 117: 123. Available online: https://www.jstor.org/stable/2997051?origin=crossref

Diniz-Filho, J.A.F., J.C. Nabout, L.M. Bini, T.N. Soares, M.P. de Campus Telles, P. de Marco, &

R.G. Collevatti. 2009. Niche modelling and landscape genetics of Caryocar brasiliense ("Pequi" tree: Caryocaraceae) in Brazilian Cerrado: An integrative approach for evaluating central-peripheral population patterns. Tree Genetics and Genomes 5: 617–627.

Ehrendorfer, F., W. Morawetz, & J. Dawe. 1984. The neotropical angiosperm families Brunelliaceae and Caryocaraceae: First karyosystematical data and affinities. Plant Systematics and Evolution 145: 183–191. Available online: http://link.springer.com/10.1007/BF00983947

Françoso, R., A. de C. Guaraldo, M. Prada, A.O. Paiva, E.H. Mota, & J.R.R. Pinto. 2014. Fenologia e produção de frutos de Caryocar brasiliense Cambess. E Enterolobium gummiferum (Mart.) J.F.Macbr. em diferentes regimes de queima. Revista Árvore 38: 579–590. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-67622014000400001&lng=pt&tlng=pt

Gonçalves, G.A.S., E.V. de B. Vilas Boas, J.V. de Resende, A.L. de L. Machado, & B.M. Vilas Boas. 2011. Qualidade dos frutos do pequizeiro submetidos a diferentes tempos de cozimento. Ciência e Agrotecnologia 35: 377–385. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-70542011000200020&lng=pt&tlng=pt

Gribel, R., & J.D. Hay. 1993. Pollination ecology of Caryocar brasiliense (Caryocaraceae) in Central Brazil cerrado vegetation. Journal of Tropical Ecology 9: 199–211. Available online: http://www.journals.cambridge.org/abstract_S0266467400007173

Grzebieluckas, C., A.C. Bornia, L.M. de S. Campos, & P.M. Selig. 2010. Evaluation of the opportunity cost for the conservation of the Cerrado in the production of pequi: A study in Mato Grosso. Custos e Agronegocio 6: 104–120.

Guimarães, P.R., M. Galetti, & P. Jordano. 2008. Seed Dispersal Anachronisms: Rethinking the Fruits Extinct Megafauna Ate. (D. M. Hansen, Ed.)PLoS ONE 3: e1745. Available online: https://dx.plos.org/10.1371/journal.pone.0001745

Leite, G.L.D., R. Von dos S. Veloso, J.C. Zanuncio, L.A. Fernandes, & C.I.M. Almeida. 2006. Phenology of Caryocar brasiliense in the Brazilian cerrado region. Forest Ecology and Management 236: 286–294. Available online: https://linkinghub.elsevier.com/retrieve/pii/S0378112706009017

Londe, L.N., C. Ueira-Vieira, W.E. Kerr, & A.M. Bonetti. 2010. Characterization of DNA polymorphisms in Caryocar brasiliense (Camb.) in populations with and without thorn at the

endocarp by RAPD markers. Anais da Academia Brasileira de Ciencias 82: 779–789.

Maciel, T.C.M., C.A. Marco, E.E. Silva, T.I. Da Silva, H.R. Dos Santos, S.D.P. Freitas Júnior, F.D. de O. Alcantara, & M.M. Chaves. 2018. Pequi (Caryocar coriaceum Wittm.) extrativism: situation and perspectives for its sustainability in Cariri Cearense, Brazil. Acta Agronómica 67: 238–245. Available online: https://revistas.unal.edu.co/index.php/acta_agronomica/article/view/62848

Melo-Júnior, Afrânio Farias de, D. de Carvalho, J.S.R. Póvoa, & E. Bearzoti. 2004. Genetic structure of natural populations of pequizeiro (Caryocar brasiliense Camb .). Scientia 1: 56–65.

Melo-Júnior, Afranio Farias de, D. de Carvalho, F.A. Vieira, & D.A. ce Oliveira. 2012. Spatial genetic structure in natural populations of Caryocar brasiliense Camb. (Caryocareceae) in the North of Minas Gerais, Brazil. Biochemical Systematics and Ecology 43: 205–209. Available online: http://dx.doi.org/10.1016/j.bse.2012.02.005

Miranda-Vilela, Ana L., L.C.S. Pereira, C.A. Gonçalves, & C.K. Grisolia. 2009. Pequi fruit (Caryocar brasiliense Camb.) pulp oil reduces exercise-induced inflammatory markers and blood pressure of male and female runners. Nutrition Research 29: 850–858. Available online: http://dx.doi.org/10.1016/j.nutres.2009.10.022

Miranda-Vilela, Ana Luisa, C.K. Grisolia, J.P.F. Longo, R.C.A. Peixoto, M.C. de Almeida, L.C.P. Barbosa, M.M. Roll, F.A. Portilho, L.L.C. Estevanato, A.L. Bocca, Sô.N. Báo, & Z.G.M. Lacava. 2014. Oil rich in carotenoids instead of vitamins C and E as a better option to reduce doxorubicin-induced damage to normal cells of Ehrlich tumor-bearing mice: hematological, toxicological and histopathological evaluations. The Journal of Nutritional Biochemistry 25: 1161–1176. Available online: http://dx.doi.org/10.1016/j.jnutbio.2014.06.005

Miranda-Vilela, Ana Luisa, R.C.A. Peixoto, J.P.F. Longo, D. de O.S. e Cintra, F.A. Portilho, K.L.C. Miranda, P.P.C. Sartoratto, S.N. Báo, R.B. de Azevedo, & Z.G.M. Lacava. 2013. Dextran-Functionalized Magnetic Fluid Mediating Magnetohyperthermia Combined with Preventive Antioxidant Pequi-Oil Supplementation: Potential Use Against Cancer. Journal of Biomedical Nanotechnology 9: 1261–1271. Available online: http://openurl.ingenta.com/content/xref?genre=article&issn=1550-7033&volume=9&issue=7&spage=1261

Miranda-Vilela, Ana Luisa, I.F. Ribeiro, & C.K. Grisolia. 2016. Association between interleukin 6 -174 G/C promoter gene polymorphism and runners' responses to the dietary ingestion of

antioxidant supplementation based on pequi (Caryocar brasiliense Camb.) oil: a before-after study. Genetics and Molecular Biology 39: 554–566. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-47572016000400554&lng=en&tlng=en

Monteiro, S.S., C. Copetti, G. Nogara, F.M. Dalla Nora, R.C. Prestes, & C.S. da Rosa. 2014. Natural antioxidant from pequi (Caryocar brasiliense Camb.) peel in the production of sausage. International Food Research Journal 21: 1963–1970.

Moura, N.F., L.J. Chaves, R.V. Naves, A.V. De Aguiar, & G.D.R. Sobierajski. 2013. Variabilidade entre procedências e progênies de Pequizero (Caryocar brasiliense Camb.). Scientia Forestalis/Forest Sciences 41: 103–112.

Oliveira, C.C., C.V. Oliveira, J. Grigoletto, L.R. Ribeiro, V.R. Funck, L. Meier, M.R. Fighera, L.F.F. Royes, A.F. Furian, I.R.A. Menezes, & M.S. Oliveira. 2017. Anticonvulsant activity of Caryocar coriaceum Wittm. fixed pulp oil against pentylenetetrazol-induced seizures. Neurological Research 39: 667–674. Available online: https://doi.org/10.1080/01616412.2017.1324380

Oliveira, L.M. de, T.S. de Oliveira, R.M. da Costa, J.L.R. Martins, C.S. de Freitas, E. de S. Gil, E.A. Costa, R. de C.A.T. Passaglia, B.G. Vaz, F.P. Filgueira, & P.C. Ghedini. 2018. Caryocar brasiliense induces vasorelaxation through endothelial Ca2+/calmodulin and PI3K/Akt/eNOS-dependent signaling pathways in rats. Revista Brasileira de Farmacognosia 28: 678–685. Available online: https://doi.org/10.1016/j.bjp.2018.07.007

Pinto, L.C.L., L.M.O. Morais, A.Q. Guimarães, E.D. Almada, P.M. Barbosa, & M.A. Drumond. 2016. Traditional knowledge and uses of the Caryocar brasiliense Cambess. (Pequi) by “quilombolas” of Minas Gerais, Brazil: subsidies for sustainable management. Brazilian Journal of Biology 76: 511–519. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-69842016000200511&lng=en&tlng=en

Prance, G T, & M.F. Silva. 1973. A monograph of Caryocaraceae. Pp. 1–75. *In*: Flora Neotropica. New York Botanical Garden Press: New York,.

Prance, Ghillean T. 2014. Caryocaraceae. Pp. 1–332. *In*: K. Kubitzki (ed.). Flowering Plants. Eudicots: Malpighiales (Vol. 11). Springer Berlin Heidelberg: Berlin, Heidelberg,.

Raiser, A.L., L. Ludwig, M.R. Marcilio, M.P.R. Torres, E.B. Ribeiro, C.R. Andrighetti, J.S. Agostini, & D.M.S. Valladão. 2018. Stability and potential antioxidant activity essay of pequi

oil (Caryocar brasiliense camb.) in cosmetic emulsions. Latin American Journal of Pharmacy 37: 144–151.

Ramos, B.H., K.L.F. Silva, R.R. Coimbra, D.B. Chagas, & W. de M. Ferreira. 2015. Anatomy and micromorphometry of Caryocar brasiliense leaves. Rodriguésia 66: 87–94. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2175-78602015000100087&lng=en&tlng=en

Ribeiro, D.A., L.G.S. De Oliveira, D.G. De Macêdo, I.R.A. De Menezes, J.G.M. Da Costa, M.A.P. Da Silva, S.R. Lacerda, & M.M.D.A. Souza. 2014. Promising medicinal plants for bioprospection in a Cerrado area of Chapada do Araripe, Northeastern Brazil. Journal of Ethnopharmacology 155: 1522–1533. Available online: http://dx.doi.org/10.1016/j.jep.2014.07.042

Santos, B.R., R. Paiva, R.C. Nogueira, L.M. De Oliveira, D.P.C. Da Silva, C. Martinotto, F.P. Soares, & P.D.D.O. Paiva. 2006. Micropropagação de pequizeiro (Caryocar brasiliense Camb.). Revista Brasileira de Fruticultura 28: 293–296.

Santos, A.M. dos, & D. Mitja. 2011. Pastagens Arborizadas No Projeto De Assentamento Benfica , Wooded Cattle Pasture in the Benfica Seetling Project in. Revista Arvore 35: 919–930.

Santos, P.H.R. dos, S.C.O. Giordani, B.C. Soares, F.H.L. e Silva, E.A. Esteves, & J.S.C. Fernandes. 2018. GENETIC DIVERGENCE IN POPULATIONS OF Caryocar brasiliense Camb. FROM THE PHYSICAL CHARACTERISTICS OF THE FRUITS. Revista Árvore 42: 146–153. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-67622018000100214&lng=en&tlng=en

Silva, A.C., E.M.S. Silva, & T.C. Silva. 2015. Apatite froth flotation using pequi's yellow pulp oil as collector. Proceedings of the World Congress on Mechanical, Chemical, and Material Engineering 1–7.

Silva, M.N. da. 2010. Extração de DNA genômico de tecidos foliares maduros de espécies nativas do cerrado. Revista Árvore 34: 973–978. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-67622010000600002&lng=pt&tlng=pt

Silva, R.R.V., L. Gomes, & U. Albuquerque. 2015. Plant extractivism in light of game theory: a case study in northeastern Brazil. Journal of Ethnobiology and Ethnomedicine 11: 6. Available online: http://www.ethnobiomed.com/content/11/1/6

Smith, M., & C. Fausto. 2016. Socialidade e diversidade de pequis (Caryocar brasiliense , Caryocaraceae) entre os Kuikuro do alto rio Xingu (Brasil) Sociality and diversity of pequi (Caryocar brasiliense Caryocaraceae) among the Kuikuro of the Upper Xingu river (Brazil). Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas 11: 87–113.

Sousa-Júnior, J.R., U.P. Albuquerque, & N. Peroni. 2013. Traditional Knowledge and Management of Caryocar coriaceum Wittm. (Pequi) in the Brazilian Savanna, Northeastern Brazil1. Economic Botany 67: 225–233.

Sousa, A.M.S., P.S.N. Lopes, L.M. Ribeiro, T.A. Santiago, V.R. Lacerda, & C.P.S. Martins. 2017. Germination and storage of Caryocar brasiliense seeds. Seed Science and Technology 45: 557–569. Available online: http://www.ingentaconnect.com/content/10.15258/sst.2017.45.3.18

Sousa, E.P. de, A.J. de M. Queiroz, R.M.F. de Figueirêdo, J.E.A. dos Santos, & D.M. Lemos. 2016. Thermophysical properties of the pequi pulp in different concentrations. Bioscience Journal 32: 20–28. Available online: http://www.seer.ufu.br/index.php/biosciencejournal/article/view/26815/17747

Sousa Júnior, J.R., R.G. Collevatti, E.M.F. Lins Neto, N. Peroni, & U.P. Albuquerque. 2018. Traditional management affects the phenotypic diversity of fruits with economic and cultural importance in the Brazilian Savanna. Agroforestry Systems 92: 11–21. Available online: http://link.springer.com/10.1007/s10457-016-0005-1

Suffredini, I.B., M.L.B. Paciencia, A.D. Varella, & R.N. Younes. 2007. In vitro cytotoxic activity of Brazilian plant extracts against human lung, colon and CNS solid cancers and leukemia. Fitoterapia 78: 223–226. Available online: https://linkinghub.elsevier.com/retrieve/pii/S0367326X07000354

Torres, L.R. de O., F.C. de Santana, F.L. Torres-Leal, I.L.P. de Melo, L.T. Yoshime, E.M. Matos-Neto, M.C.L. Seelaender, C.M.M. Araújo, B. Cogliati, & J. Mancini-Filho. 2016. Pequi (Caryocar brasiliense Camb.) almond oil attenuates carbon tetrachloride-induced acute hepatic injury in rats: Antioxidant and anti-inflammatory effects. Food and Chemical Toxicology 97: 205–216. Available online: http://dx.doi.org/10.1016/j.fct.2016.09.009

Vilas Boas, B.M., G.A.S. Gonçalves, J.A. Alves, J.M. Valério, T.C. Alves, L.J. Rodrigues, R.H. Piccoli, & E.V. de B. Vilas Boas. 2012. Qualidade de pequis fatiados e inteiros submetidos ao congelamento. Ciência Rural 42: 904–910. Available online: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-

84782012000500024&lng=pt&tlng=pt

Watling, J., M.P. Shock, G.Z. Mongeló, F.O. Almeida, T. Kater, P.E. De Oliveira, & E.G. Neves. 2018. Direct archaeological evidence for Southwestern Amazonia as an early plant domestication and food production centre. (J. P. Hart, Ed.)PLOS ONE 13: e0199868. Available online: https://dx.plos.org/10.1371/journal.pone.0199868

Wurdack, K.J., & C.C. Davis. 2009. Malpighiales phylogenetics: Gaining ground on one of the most recalcitrant clades in the angiosperm tree of life. American Journal of Botany 96: 1551–1570. Available online: http://doi.wiley.com/10.3732/ajb.0800207

Xi, Z., B.R. Ruhfel, H. Schaefer, A.M. Amorim, M. Sugumaran, K.J. Wurdack, P.K. Endress, M.L. Matthews, P.F. Stevens, S. Mathews, & C.C. Davis. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proceedings of the National Academy of Sciences 109: 17519–17524. Available online: http://www.pnas.org/cgi/doi/10.1073/pnas.1205818109

6. ACKNOWLEDGMENT

This work was developed in the context of the National Institutes for Science and Technology in Ecology, Evolution and Biodiversity Conservation (INCT - EECBio), supported by MCTIC/CNPq (process #465610/2014-5) and Foundation for Research Support of the State of Goiás (FAPEG), in addition to support from PPGS CAPES/FAPEG (Public Call #08/2014) and National Council for Scientific and Technological Development (CNPq) (Call MCTIC/CNPq #28/2018, process 435477/2018-8). R.N. was supported by doctoral fellowship from Coordination for the Improvement of Higher Education Personnel (CAPES). M.P.C.T. was supported by productivity fellowship from CNPq.

CAPÍTULO 2

DATA ON THE DRAFT GENOME SEQUENCE OF *Caryocar brasiliense* Camb. (CARYOCARACEAE): AN IMPORTANT GENETIC RESOURCE FROM BRAZILIAN SAVANNAS[1]

Rhewter Nunes[2]; Ariany Rosa Gonçalves[2]; Mariana Pires de Campos Telles [2,3]

---

[1] Capítulo publicado como artigo no periódico científico *Data in Brief*;

[2] Genetics & Biodiversity Laboratory (LGBio), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, GO, Brasil.
[3] School of Agrarian and Biological Sciences, Pontifical Catholic University (PUC - GO), Goiânia, Brazil.

NUNES, R.; GONÇALVES, A. R.; TELLES, M. P. C. Data on the draft genome sequence of *Caryocar brasiliense* Camb. (Caryocaraceae): an important genetic resource from Brazilian savannas, **Data in Brief**, 2019.

Data in brief 26 (2019) 104543

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data on the draft genome sequence of *Caryocar brasiliense* Camb. (Caryocaraceae): An important genetic resource from Brazilian savannas

Rhewter Nunes [a, *], Ariany Rosa Gonçalves [a], Mariana Pires de Campos Telles [a, b]

[a] *Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas - UFG, Goiânia, 74690-900, Brazil*
[b] *Escola de Ciências Agrárias e Biológicas, PUC-GO, Goiânia, Brazil*

ARTICLE INFO

Article history:
Received 23 May 2019
Received in revised form 27 August 2019
Accepted 12 September 2019
Available online 23 September 2019

Keywords:
Cerrado
Genome assembly
Genome evolution
Native plants
SSR markers

ABSTRACT

Caryocar brasiliense (Caryocaraceae) is a Neotropical tree species widely distributed in Brazilian savannas. This species is very popular in central Brazil mainly due to the use of its fruits in the local cuisine and their anti-inflammatory proprieties, and indeed it is one of the candidates, among Brazilian native plants, for fast track incorporation into cropping systems. Considering the importance of *Caryocar brasiliense*, little is known about its genetics and genomics, and determination of a reference genome sequence could improve the understanding of its evolution, as well as the development of tools for domestication. Here, we provide the first draft genome of *C. brasiliense*, the raw sequencing data and some multiplex sets of high quality microsatellite primers. Data on the genome project can be obtained from the BioProject at NCBI (https://www.ncbi.nlm.nih.gov/bioproject/?term=caryocar).

* Corresponding author.
E-mail address: rhewter@gmail.com (R. Nunes).

https://doi.org/10.1016/j.dib.2019.104543

Specifications Table

Subject area	Biology
More specific subject area	Genomics, horticultural science.
Type of data	Whole genome sequence data, genome assembly and primers for candidate microsatellites markers.
How data were acquired	High-throughput sequencing (Illumina HiSeq 2000).
Data format	Raw sequencing reads (fastq) and draft-genome (fasta).
Experimental factors	Sequencing was performed using Illumina HiSeq, and the draft genome was determined using Platanus software.
Experimental features	Sequencing was performed according to Illumina Nextera protocol for DNA-Seq.
Data source location	Agronomy School, Federal University of Goiás - Goiânia, Goiás, Brazil (16°35′49.8″S 49°16′45.4″W).
Data accessibility	The complete genome sequence of *Caryocar brasiliense* is available in the NCBI GenBank under accession number: STGP00000000. The sequencing reads used in assembly analysis are available in the NCBI SRA database under accession number: SRX5692978 (https://www.ncbi.nlm.nih.gov/sra/?term= SRX5692978).

Value of the Data

- This dataset provides the first version of a draft genome for *Caryocar brasiliense*. This is the first genome project for a species from the Caryocaraceae family and can be used as a reference in future genome projects for other species.
- This dataset can be used for comparative analyses in evolutionary studies. The draft genome can be used to identify genes, repeat regions, microsatellites and other genome elements that can describe the biology and evolution of the species.
- Primer data can be used for the development of molecular markers for domestication and breeding programs. We selected and made available some high quality multiplex microsatellite sets for genetic diversity analysis.

1. Data

The pequi (*Caryocar brasiliense* Camb.) belongs to the family Caryocaraceae (Malpighiales order) and is an important genetic resource from Brazilian savannas mainly because of the use of its fruits in local cuisine and their anti-inflammatory proprieties. We present the first draft genome of *C. brasiliense* using high-throughput DNA sequencing, the raw sequencing data used in the genome assembly analysis and a set of primers to amplify candidate microsatellite markers. The draft genome recovered 45.69% of the estimated genome size (464,365,380 bp) distributed in 55,248 contigs (Table 1). The draft genome is available at: https://www.ncbi.nlm.nih.gov/nuccore/STGP00000000.1/. The raw reads dataset was obtained from a run using Illumina HiSeq2000 equipment. A total of 293,621,819 sequencing reads of 100 base pairs each were generated. Sequencing data are available at: https://www.ncbi.nlm.nih.gov/sra/?term= SRX5692978. Additionally, 5 multiplex with 5 to 7 high-quality microsatellite primers (total of 30 pairs of primers) were designed and are available in this paper (Table 2).

2. Experimental design, materials, and methods

2.1. Total DNA sampling and sequencing

Fresh leaves were collected from a tree at Escola de Agronomia, Universidade Federal de Goiás, Goiânia, Goiás, Brazil (16°35′49.8″S 49°16′45.4″W). The total DNA was extracted from leaves using the CTAB protocol [1]. The quality of DNA was determined by a Nanodrop device, and the quantity was measured by a Qbit and 1% agarose gel. The sample was sent to Centro de Genômica Funcional ESALQ-

Table 1
Genome assembly statistics of the draft genome of *Caryocar brasiliense*.

Metric	Value
Number of contigs	55,248
Number of contigs ≥1000 bp	43,286
Total length	212,172,521
Largest contig	64,707
Shortest contig	500
N50	6005
N75	3615
L50	10,532
L75	21,784
GC%	34.84

USP core facility for sequencing. An Illumina paired-end 2 × 100 bp library was constructed and forwarded for sequencing using an Illumina HiSeq2000 platform.

2.2. Sequencing quality control and assembly

Raw reads were evaluated for base quality sequencing and sequencing adapter presence using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality control was performed using Trimmomatic software v0.39 [2] with the options ILLUMINACLIP: TruSeq3-PE.fa:2:

Table 2
Multiplex microsatellite primers designed for *Caryocar brasiliense*.

Multiplex_ID	Primer_ID	SSR_Motif	Primer_Foward_5′-3′	Primer_reverse_5′_3′	Ta	PCR_Frag_len
1	Cbr_NGS_SSR1	TATG	gctacttccagtcactagacttgt	cacaactgctaccatgttcgac	62	349
1	Cbr_NGS_SSR2	CATA	acccgccttctccagtgaata	tcctcgagttttacagcggtat	60	164
1	Cbr_NGS_SSR3	CT	ctctctttgcgggatatctcaaga	ccatgacagtccagcccaata	61	224
1	Cbr_NGS_SSR4	CT	actctgccgacagctgaattta	aaaggcaacacagcagatcattaa	60	102
1	Cbr_NGS_SSR5	AG	gtggaaatgcataaactgtatgcct	cgatagctgctcttgccaagt	62	584
1	Cbr_NGS_SSR6	TC	gcttctgcaaaatcataggcaaca	agtggtaattcacgctggtaattta	60	425
1	Cbr_NGS_SSR7	TTC	gccattctcaattttccagtggac	gtgtgtgttgtaaacattcaaggat	60	493
2	Cbr_NGS_SSR8	AGG	aataagatgccattgcggtgtt	tgaccgactctttcttattgggaa	60	157
2	Cbr_NGS_SSR9	TC	tacataaattgtcttcagcccatgt	agcctgctcgattaagtgaaca	60	278
2	Cbr_NGS_SSR10	GCA	agagtccttgtgacgaatcagatt	ctcatccgagaacttatgcagc	60	218
2	Cbr_NGS_SSR11	GAT	gccatcagcgaacagttctct	caacaaattacctgctccgagtt	61	372
2	Cbr_NGS_SSR12	TTC	gagttttgatgcttaagccatgac	gccttaccagagtctgcaagt	61	434
2	Cbr_NGS_SSR13	GGT	ccactgacttattcaatttctcgac	ggaccctcaacaggacctattt	60	513
3	Cbr_NGS_SSR14	AG	gaactcttttccctacagatcagaa	catttcaggttgagtagcttgtca	60	270
3	Cbr_NGS_SSR15	GCT	ggacgccatttcacaagattga	ccctgctgtcaacaggattct	61	132
3	Cbr_NGS_SSR16	CTT	aggatgcctttccaaagacgt	ttttacagcaacatttgtgagactc	60	331
3	Cbr_NGS_SSR17	CAA	ttaatgatctggggtcacatcctt	gtgggggcaatggacctaatat	60	195
3	Cbr_NGS_SSR18	GTT	ggagatcagaccaagcattgct	tgcatcattttggcgactacaat	61	495
3	Cbr_NGS_SSR19	TTC	gaggctgcattaagcatggaaa	aagacaaaagagtggatttcccac	61	402
4	Cbr_NGS_SSR20	GAA	aaaactggtagaagatgcagtcaa	gattagaatgtgcaaaattggcagt	60	312
4	Cbr_NGS_SSR21	CTT	aacggggtcccatcgtatctt	gacacctgttaagcaagaacatgt	62	251
4	Cbr_NGS_SSR22	CTT	cggtatatggaagcgtacttcac	tctgcactcgcaagtccaata	60	176
4	Cbr_NGS_SSR23	GTT	gcttttgttgtggagccaaattaca	cgcgaaattcctcatgttcaga	60	109
4	Cbr_NGS_SSR24	GTT	gtcattaacctgacaccattgct	tctactgctatgttcggagcatatt	61	392
5	Cbr_NGS_SSR25	GA	tattcaggcgtggcaccaata	tggctcaaaactttgcatactgat	61	258
5	Cbr_NGS_SSR26	GA	ctgcttcagttcggagaccaa	atctacttccaaagacatagtgtgc	61	332
5	Cbr_NGS_SSR27	GA	cgtcaaatcttccaacagctga	catgtttcattgaagggccatcat	60	180
5	Cbr_NGS_SSR28	CT	aggtgatgtgaccttccaagc	agaatggggattcgtgttctagtt	61	447
5	Cbr_NGS_SSR29	GA	ctagcagtgcttcgtcaaaactt	ttattcagtgacccggttatggat	60	111
5	Cbr_NGS_SSR30	TC	gttcagcaaacattctgctaagtc	ttgggaacgtaaagatcaatttcct	60	508

30:10 and SLIDEWINDOW: 4:30, which required at least a mean Phred score of 30 for every four bases. The best k-mer value was estimated using Kmergenie software [3]. The *de novo* assembly was performed using Platanus (PLATform for Assembling NUcleotide Sequences) software v1.2.4 [4].

2.3. Microsatellite identification and primer design

The microsatellite regions were identified in the genome using QDD software [5]. The program marks the primers for microsatellite regions that occur in the context of transposable elements. This allows the selection of the best primer pairs for the molecular marker test as it minimizes the occurrence of null alleles due to primer annealing problems. We used only contigs larger than 10 Kb in the microsatellite analysis. After identification of the microsatellite regions, we applied a rigorous filter to choose the best sets of primers for molecular marker tests. Among the 120,858 pairs of primers designed for 6885 identified microsatellite regions we applied the following filters: i) primers with a size between 20 and 24 base pairs; ii) PCR product size between 150 and 460 base pairs; iii) not including a region formed only by adenine and thymine bases; iv) at least 16 dinucleotide, 6 trinucleotide, 6 tetranucleotide and 4 pentanucleotide repeats and v) the difference in annealing temperature between the primers is less than 2 °C. For the resulting set of primers, the best pair for each microsatellite region was chosen based on the greatest possible distance between target regions and primers. We used FastPCR software to generate the multiplex sets [6]. The final set of primers we recommend for testing as molecular markers correspond to 30 microsatellite regions distributed in a set of 5 PCR multiplex.

Acknowledgments

This work was developed in the context of the National Institutes for Science and Technology in Ecology, Evolution and Biodiversity Conservation (INCT - EECBio), supported by MCTIC/CNPq (process #465610/2014-5) and the Foundation for Research Support of the State of Goiás (FAPEG), in addition to support from PPGS CAPES/FAPEG (Public Call #08/2014) and the National Council for Scientific and Technological Development (CNPq) (Call MCTIC/CNPq #28/2018, process 435477/2018-8). R.N. and A.R.G. were supported by doctoral fellowships from Coordination for the Improvement of Higher Education Personnel (CAPES). M.P.C.T. was supported by productivity fellowships from CNPq.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Doyle, J.J. Doyle, Doyle&Doyle_Focus_1990_CTAB.pdf, Focus 12 (1990) 13–15.
[2] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.
[3] R. Chikhi, P. Medvedev, Informed and automated k-mer size selection for genome assembly, Bioinformatics 30 (2014) 31–37, https://doi.org/10.1093/bioinformatics/btt310.
[4] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, T. Itoh, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads, Genome Res. 24 (2014) 1384–1395, https://doi.org/10.1101/gr.170720.113.
[5] E. Meglécz, C. Costedoat, V. Dubut, A. Gilles, T. Malausa, N. Pech, J.F. Martin, QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects, Bioinformatics 26 (2009) 403–404, https://doi.org/10.1093/bioinformatics/btp670.
[6] R. Kalendar, D. Lee, A.H. Schulman, FastPCR software for PCR, in silico PCR, and oligonucleotide assembly and analysis, in: S. Valla, R. Lale (Eds.), DNA Cloning and Assembly Methods. Methods in Molecular Biology (Methods and Protocols), vol. 1116, Humana Press, Totowa, NJ, 2014.

CAPÍTULO 3

COMPLETE CHLOROPLAST GENOME SEQUENCE OF *Caryocar brasiliense* Camb. (CARYOCARACEAE) AND COMPARATIVE ANALYSIS BRINGS NEW INSIGHTS INTO THE PLASTOME EVOLUTION OF MALPIGHIALES[1]

Rhewter Nunes[2]; Ueric José Borges de Souza[2]; Cintia Pelegrineti Targueta[2]; Rafael Barbosa Pinto[2]; Thannya Nascimento Soares[2]; José Alexandre Felizola Diniz-Filho[3]; Mariana Pires de Campos Telles [2,4]

---

[1] Capítulo recomendado para publicação no periódico científico *Genetics and Molecular Biology*;

[2] Genetics & Biodiversity Laboratory (LGBio), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, GO, Brasil.
[3] Laboratory of Theoretical Ecology and Synthesis (LETS), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, 74690-900, Brazil
[4] School of Agrarian and Biological Sciences, Pontifical Catholic University (PUC - GO), Goiânia, Brazil.

Title:

Complete chloroplast genome sequence of *Caryocar brasiliense* Camb. (Caryocaraceae) and comparative analysis brings new insights into the plastome evolution of Malpighiales

Authors:
Rhewter Nunes [1,*], Ueric José Borges de Souza [1], Cintia Pelegrineti Targueta [1], Rafael Barbosa Pinto [1], Thannya Nascimento Soares [1], José Alexandre Felizola Diniz-Filho [2] and Mariana Pires de Campos Telles [1,3]

Affiliations:

[1] Genetics & Biodiversity Laboratory (LGBio), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, Goiás, Brazil, 74690-900;
[2] Laboratory of Theoretical Ecology and Synthesis (LETS), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, Goiás, Brazil, 74690-900;
[3] School of Agrarian and Biological Sciences, Pontifical Catholic University (PUC - GO), Goiânia, Goiás, Brazil, 74885-460.
[*] Correspondence: rhewter@gmail.com.

Running title:

Plastome of *Caryocar brasiliense*

Keywords:
Cerrado; genomics; molecular evolution; organellar genome; plastome evolution.

Correspondence:

Rhewter Nunes

Genetics & Biodiversity Laboratory (LGBio), Institute of Biological Sciences - Federal University of Goiás (UFG), Goiânia, Goiás, Brazil, 74690-900. Cel. +55 62 99904-5107. E-mail: rhewter@gmail.com

Abstract:

Caryocar brasiliense (Caryocaraceae) is a Neotropical tree species widely distributed in Brazilian Savannas. This species is very popular in central Brazil mainly by the use of its fruits in the local cuisine, and indeed it is one of the candidates, among Brazilian native plants, for fast track incorporation into cropping systems. Here we sequenced the complete chloroplast genome of *C. brasiliense* and used the data to access its genomic resources using high-throughput sequencing. The chloroplast exhibits a genome length of 165,793 bp and the typical angiosperm quadripartite structure with two copies of an inverted repeat sequence (IRa and IRb) of 34,902 bp each, separating a small single copy (SSC) region of 11,852 bp and a large single copy (LSC) region of 84,137 bp. The annotation analysis identified 136 genes being 87 protein-coding, eight rRNA and 37 tRNA genes. We identified 49 repetitive DNA elements and 85 microsatellites. A bayesian phylogenetic analysis helped to understand previously unresolved relationships in Malpighiales, placing Caryocaraceae as a separated group in the order, with high supported nodes. This study synthetizes valuable information for further studies allowing a better understanding of evolutionary patterns in the group and providing resources for future breeding programs.

A group not yet explored in terms of genomic approaches in the order Malpighiales is the family Caryocaraceae. This family is one of the poorly resolved groups within Malpighiales, forming a polytomy with some other families such as Malpighiaceae and Chrysobalanaceae, groups for which we have fully sequenced chloroplast genomes for the species (Xi et al. 2012; APG IV et al. 2016). The main representatives of the Caryocaraceae family are the species of the genus *Caryocar* L., especially *Caryocar brasiliense* Camb. This species is a Neotropical tree much valued in Brazilian cuisine, appreciated as a nutritious resource for bats (Gribel and Hay 1993) and widely known in folk culture as a symbol of Brazilian savannas (or Cerrado) (Gribel and Hay 1993; De Araujo 1995). Also, the fruit pulp of *C. brasiliense* is rich in unsaturated fatty acids, vitamins and phenolic acids, as well carotenoids such as violaxanthin, lutein and zeaxanthin (Castro et al. 2008; Mariano et al. 2009). Because of all these characteristics, *C. brasiliense* is one of the main native Cerrado species that are candidates for incorporation into cropping systems (Leite et al. 2006; Tunholi et al. 2013).

Despite the importance of *Caryocar brasiliense*, until now (January, 2019) no genomic resources were developed to this species. A few studies have used a microsatellites markers or short sequences to evaluate genetic diversity patterns, showing that the natural populations of *C. brasiliense* have relatively low genetic structure and a great deal of genetic and phylogeographic diversity (e.g. Diniz-Filho et al. 2009; Collevatti et al. 2011). Thus, here we sequenced the complete chloroplast genome of *C. brasiliense* and used the data to access its genomic resources using high-throughput sequencing. We generate information of chloroplast genome sequence, gene composition and organization, and repeat sequences, using part of these data to reconstruct a phylogenetic tree for Malpighiales order and analyze relationships of *C. brasiliense* within this group.

The total DNA of fresh leaves of *Caryocar brasiliense* was extracted from leaves using CTAB protocol. The sample was sequenced in an Illumina HiSeq2000 platform in paired-end 2x100 pb mode. Raw reads were evaluated for base quality sequencing and sequencing adapters

presence using FastQC software (Andrews 2010). The quality control was performed using Trimmomatic (Bolger et al. 2014) software. The high-quality reads were taken for a *de novo* chloroplast genome assembly in NOVOPlasty v.2.7.1 software (Dierckxsens et al. 2017). We performed a gene annotation of *Caryocar brasiliense* chloroplast genome using CHLOROBOX GeSeq (Tillich et al. 2017) and DOGMA - Dual Organellar GenoMe Annotator (Wyman et al. 2004) software. Simple sequence repeats (SSR) or microsatellite regions were predicted in *C. brasiliense* chloroplast genome using IMEx - Imperfect Microsatellite Extractor (Mudunuri and Nagarajaram 2007) and repeat sequence elements using REPuter software (Kurtz 2002). Also, we performed a bayesian phylogenetic analysis among some Malpighiales species using 76 protein-coding gene sequences. For a more detailed description of the methods see supplementary material.

The complete sequence of *Caryocar brasiliense* chloroplast genome was deposited into GenBank (accession number: MK726375) with high mean genome coverage (715X). The plastome of *C. brasiliense* exhibited a total length of 165,793 bp and typical quadripartite division, which is also observed in other flowering plants (Figure 1; Figure S1). The genome comprehended a Large Single Copy (LSC), a Small Single Copy (SSC) and a pair of Inverted Repeats (IRa and IRb). These regions had 84,137 bp, 11,852 bp and 34,902 bp, respectively. Besides, we observed a GC content of 36.7%. Inverted repeats regions exhibited the greatest GC content value with 39.6%, followed by LSC (35.0%) and SSC (31.5%). Inside inverted repeat regions, the GC content was bigger where rRNAs were predicted.

We compared the structural features of *C. brasiliense* chloroplast genome with other nine chloroplast genomes from nine other families in Malpighiales order. Compared to then, *C. brasiliense* had the largest genome size with large Inverted Repeat regions, but one of the smallest Small Single Copies region, along with *Linum usitatissimum* L. (de Souza et al. 2017). However, *C. brasiliense* show similar values of the size of LSC region when compared to other species under analysis. This may be explained by the fact that many of the photosynthetic genes are present in

this region and so, they are important for the persistence of these species resulting in fewer contraction / expansion events in that region of the chloroplast genome among species.

We found 115 different genes in the genome, of which 77 were protein-coding genes, four ribosomal RNAs, 30 transfer RNAs and four pseudogenes (Table S1). Also, 10 protein-coding genes, four rRNAs and seven tRNAs were observed as duplicated. Three of these tRNA genes, trnA-UGC, trnM-CAU and trnR-ACG, had more than one copy in the genome (each gene appeared 4 times). So, considering the duplicated genes, the *C. brasiliense* chloroplast genome had a total of 136 genes (87 protein-coding genes, four pseudogenes, eight rRNAs and 37 tRNAs). We also observed 10 genes that had introns. In a general view, *C. brasiliense* has a relatively conserved number of genes when compared to other Malpighiales species, specially related to rRNA, tRNA and photosynthesis related genes. The gene features were very similar to *Byrsonima coccolobifolia* Kunth (Menezes et al. 2018), another Malpighiales species (Table S2).

The pseudogenes predicted in *C. brasiliense* chloroplast genome were ndhH, psaA, psaB and psbA. All of them have a copy of the gene in the complete form and are supposed to be involved in structural changes in the chloroplast genome of *C. brasiliense*. Such as ndhH pseudogene was related to the process of duplication of Inverted Repeat region, the others pseudogenes can be related to inversion events in Large Single Copy region generating variation in the order of the plastid genes when compared to others Malpighiales chloroplast genomes. Such inversion events were also observed in chloroplast genome of *Passiflora edulis* Sims, another Malpighiales species (Cauz-Santos et al. 2017), which can be an indicative of a non-conservative gene collinearity related to all members of this order.

The comparative analysis with ten Malpighiales species revealed a high conserved pattern of variation along the genome (very variable regions between *C. brasiliense* and *Jatropha curcas* are also very variable in other species compared to *J. curcas*) with protein-coding genes highly conserved, as well as intergenic regions with more variation (Figure S2). *Passiflora edulis* was the

species with higher divergent regions in comparison to *Jatropha curcas* L. than other species, presenting more regions with similarity below to 50%.

Compared to other species under analysis, *Caryocar brasiliense* displays one of the smallest Small Single Copy (SSC) regions and the greatest Inverted Repeat (IR) regions. This indicates that the size of these regions evolves differently between different species of Malpighiales. (Mower et al. 2015). We performed an IR boundaries comparison analyses to investigate which genes were present in the sites that separate the chloroplast regions (Figure 2). The most common gene that flanks the region between the IR and SSC is ycf1 presented in seven of the ten analyzed species. Commonly, the a ycf1 gene and a pseudogene of ycf1 were present in IR/SSC boundaries in Malpighiales such as *Byrsonima coccolobifolia* (Menezes et al. 2018), *Chrysobalanus icaco* L. (Bardon et al. 2016), *Erythroxylum novogranatense* (D. Morris) Hieron., *Garcinia mangostana* L. (Jo et al. 2017), *Manihot esculenta* Crantz (Daniell et al. 2008), *Populus tremula* L. (Kersten et al. 2016) and *Viola seoulensis* Nakai (Cheon et al. 2017).

Three of the species under analysis did not show ycf1 as flanking gene of IR/SSC region: *Caryocar brasiliense*, *Passiflora edulis* (Cauz-Santos et al. 2017) and *Linum usitatissimum* L. (de Souza et al. 2017). *Caryocar brasiliense* has the bounder between IR/SSC flanked by ndhH gene and ndhH pseudogene, whereas in *P. edulis* rps15 and intergenic region ycf1-ndhF are found. For *L. usitatissimum,* the ndhA gene and the intergenic region ndhA-ndhF were observed. The IR / SSC boundary region presents high collinearity when comparing different genomes. The order of the genes is conserved as: ndhA, ndhH, rps15 and ycf1 for all species under analysis. While the bounder of IR/SSC for the major species include ycf1, in some Malpighiales this site was displaced with the donation of genomic segments from the SSC region to the IR regions. In addition to colinear genes evidence, this pattern of expansion of IR regions with contraction of SSC region was also supported by the results of regions length indicating that species with no ycf1 flanking IR/SSC bounder had a smaller SSC region (Table S2).

A total of 85 perfect microsatellites (SSRs) were identified in *Caryocar brasiliense* chloroplast genome sequence (Figure S3). About the repeat motif, we found 52 mononucleotide, 11 dinucleotide, five trinucleotide, 12 tetranucleotide, three pentanucleotide and two hexanucleotide. The number of continuous repeats (motif iterations) ranged from three to 16 times (Table S3). The chloroplast region with the major number of SSRs were Large Single Copy (56.47%), followed by Inverted Repeats (29.41%) and Small Single Copy (14.12%). SSRs are important regions because they can serve as molecular markers in studies of genetic diversity, phylogeny and phylogeography for Brazilian Savannas species (Rabelo et al. 2011; Soares et al. 2012; Telles et al. 2013). The identified SSR regions can be used in molecular marker development testing for genetic diversity studies with *C. brasiliense* and nearby species (Table S4).

We also identified repeat sequences in *C. brasiliense* chloroplast genome and other 10 species from Malpighiales order (chloroplast genome subset described in Material and Methods) (Table S5). We observed a total of 49 repeats in *C. brasiliense*. Considering the repeat type, we found 18 forward, 30 palindromic and one reverse. No repeats of the type complement were observed for *C. brasiliense*. Complement repeats occur only in three of the 10 species in analysis: one in *Chrysobalanus icaco*, two in *Erythroxylum novogranatense* and one in *Viola seoulensis*. Other three species, *Linum usitatissimum*, *Manihot esculenta* and *Passiflora edulis*, present only forward and palindromic repeats. Repeat analysis revealed a highly conserved total number of repeats within the Malpighiales order although the type of repeats varied in each species under analysis.

We performed phylogenetic analysis sampling 52 representatives from all families in Malpighiales order that had their chloroplast genomes sequenced using its protein-coding gene sequences (Table S6). Additionally, we also retrieved chloroplast gene sequences from *Anthodiscus peruanus* Baill. (Caryocaraceae) and *Putranjiva roxburghii* Wall. (Putranjivaceae). This analysis resulted in a phylogenetic tree with high supported values for the nodes considering Bayesian posterior probability given to each one (Figure 3). As expected, all analyzed species (including *C.*

brasiliense) fell within the clades that represent their respective botanical families, validating the chloroplast sequences obtained in this work.

The phylogeny currently accepted for Malpighiales display a polytomy involving Caryocaraceae, Putranjivoids, Malpighioids and Chrysobalanoids species (APG II et al. 2003; Xi et al. 2012). Here, for the first time, a highly supported phylogenetic tree placed Caryocaraceae as a sister clade in relation to the other families within Malpighiales (Figure 3). Also, this result provides evidences that reinforce the non-clustering among Caryocaraceae, Linaceae and Erythroxylaceae in the same clade as discussed previously (Soltis et al. 2007). Current phylogeny of this group was performed using few genes and the new genomic resources produced by this work helps for a better understanding of the phylogenetic relationships in Malpighiales order (APG II et al. 2003; APG III et al. 2009; Xi et al. 2012; APG IV et al. 2016). These new sources helped in the solution of uncertain clades and demonstrated how the use of high-throughput sequencing technologies can increase the accuracy of phylogenetic analysis. Moreover, these data increase the genetic and genomic resources available for Malpighiales offering the first complete genome sequence and content of a chloroplast in the Caryocaraceae family.

Acknowledgements:

The authors thanks to Ariany Rosa Gonçalves for the help in DNA extraction process. This work was developed in the context of the National Institutes for Science and Technology in Ecology, Evolution and Biodiversity Conservation (INCT - EECBio), supported by MCTIC/CNPq (process #465610/2014-5) and Foundation for Research Support of the State of Goiás (FAPEG), in addition to support from PPGS CAPES/FAPEG (Public Call #08/2014) and National Council for Scientific and Technological Development (CNPq) (Call MCTIC/CNPq #28/2018, process 435477/2018-8). R.N. and U.J.B.S. were supported by doctoral fellowships from Coordination for

the Improvement of Higher Education Personnel (CAPES). T.N.S., J.A.F.D.F. and M.P.C.T. were supported by productivity fellowships from CNPq.

References:

APG II (Angiosperm Phylogeny Group), Bremer B, Bremer K, Chase MW, Reveal JL, Soltis DE, Soltis PS and Stevens PF (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot J Linn Soc 141:399–436. doi: 10.1046/j.1095-8339.2003.t01-1-00158.x

APG III (Angiosperm Phylogeny Group), Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, Soltis DE, Soltis PS and Stevens PF (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121. doi: 10.1111/j.1095-8339.2009.00996.x

APG IV (Angiosperm Phylogeny Group), Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS et al. (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181:1–20. doi: 10.1111/boj.12385

Bardon L, Sothers C, Prance GT, Malé PJG, Xi Z, Davis CC, Murienne J, García-Villacorta R, Coissac E, Lavergne S et al. (2016) Unraveling the biogeographical history of chrysobalanaceae from plastid genomes. Am J Bot 103:1089–1102. doi: 10.3732/ajb.1500463

Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi: 10.1093/bioinformatics/btu170

Castro AJ, Grisolia CK, de Araújo BC, Dias CD, Dutra ES and Nepomuceno JC (2008) Recombinogenic effects of the aqueous extract of pulp from pequi fruit (Caryocar brasiliense) on somatic cells of Drosophila melanogaster. Genet Mol Res 7:1375–1383. doi: 10.4238/vol7-

4gmr515

Cauz-Santos LA, Munhoz CF, Rodde N, Cauet S, Santos AA, Penha HA, Dornelas MC, Varani AM, Oliveira GCX, Bergès H et al. (2017) The Chloroplast Genome of Passiflora edulis (Passifloraceae) Assembled from Long Sequence Reads: Structural Organization and Phylogenomic Studies in Malpighiales. Front Plant Sci 8:1–17. doi: 10.3389/fpls.2017.00334

Cheon K-S, Yang J-C, Kim K-A, Jang S-K and Yoo K-O (2017) The first complete chloroplast genome sequence from Violaceae (Viola seoulensis). Mitochondrial DNA Part A 28:67–68. doi: 10.3109/19401736.2015.1110801

Collevatti RG, Nabout JC and Diniz-Filho JAF (2011) Range shift and loss of genetic diversity under climate change in Caryocar brasiliense, a Neotropical tree species. Tree Genet Genomes 7:1237–1247. doi: 10.1007/s11295-011-0409-z

Daniell H, Wurdack KJ, Kanagaraj A, Lee S, Saski C and Jansen RK (2008) The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. Theor Appl Genet 116:723–737. doi: 10.1007/s00122-007-0706-y

Araujo FD (1995) A review of caryocar brasiliense (caryocaraceae)—an economically valuable species of the central brazilian cerrados. Econ Bot 49:40–48. doi: 10.1007/BF02862276

Souza EM, Guerra MP, Vieira L do N, Nodari RO, Rogalski M, de Oliveira Pedrosa F, Pacheco TG, de Santana Lopes A and Santos KG dos (2017) The Linum usitatissimum L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. Plant Cell Rep 37:307–328. doi: 10.1007/s00299-017-2231-z

Dierckxsens N, Mardulyn P and Smits G (2017) NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. doi: 10.1093/nar/gkw955

Diniz-Filho JAF, Nabout JC, Bini LM, Soares TN, de Campus Telles MP, de Marco P and

Collevatti RG (2009) Niche modelling and landscape genetics of Caryocar brasiliense ("Pequi" tree: Caryocaraceae) in Brazilian Cerrado: An integrative approach for evaluating central-peripheral population patterns. Tree Genet Genomes 5:617–627. doi: 10.1007/s11295-009-0214-0

Gribel R and Hay JD (1993) Pollination ecology of Caryocar brasiliense (Caryocaraceae) in Central Brazil cerrado vegetation. J Trop Ecol 9:199–211. doi: 10.1017/S0266467400007173

Jo S, Kim HW, Kim YK, Sohn JY, Cheon SH and Kim KJ (2017) The complete plastome of tropical fruit Garcinia mangostana (Clusiaceae). Mitochondrial DNA Part B Resour 2:722–724. doi: 10.1080/23802359.2017.1390406

Kersten B, Faivre Rampant P, Mader M, Le Paslier M-C, Bounon R, Berard A, Vettori C, Schroeder H, Leplé J-C and Fladung M (2016) Genome Sequences of Populus tremula Chloroplast and Mitochondrion: Implications for Holistic Poplar Breeding. PLoS One 11:e0147209. doi: 10.1371/journal.pone.0147209

Kurtz S (2002) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642. doi: 10.1093/nar/29.22.4633

Leite GLD, Von dos S. Veloso R, Zanuncio JC, Fernandes LA and Almeida CIM (2006) Phenology of Caryocar brasiliense in the Brazilian cerrado region. For Ecol Manage 236:286–294. doi: 10.1016/j.foreco.2006.09.013

Mariano RG de B, Couri S and Freitas SP (2009) Enzymatic technology to improve oil extraction from Caryocar brasiliense camb. (Pequi) Pulp. Rev Bras Frutic 31:637–643. doi: 10.1590/S0100-29452009000300003

Menezes APA, Resende-Moreira LC, Buzatti RSO, Nazareno AG, Carlsen M, Lobo FP, Kalapothakis E and Lovato MB (2018) Chloroplast genomes of Byrsonima species (Malpighiaceae): comparative analysis and screening of high divergence sequences. Sci Rep 8:2210. doi: 10.1038/s41598-018-20189-4

Mower JP, Zhu A, Guo W, Fan W and Gupta S (2015) Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol 209:1747–1756. doi: 10.1111/nph.13743

Mudunuri SB and Nagarajaram HA (2007) IMEx: Imperfect microsatellite extractor. Bioinformatics 23:1181–1187. doi: 10.1093/bioinformatics/btm097

Rabelo SG, Teixeira CF, Telles MPC and Collevatti RG (2011) Development and characterization of microsatellite markers for Lychnophora ericoides, an endangered Cerrado shrub species. Conserv Genet Resour 3:741–743. doi: 10.1007/s12686-011-9447-y

Soares TN, Melo DB, Resende LV, Vianello RP, Chaves LJ, Collevatti RG and Telles MP de C (2012) Development of microsatellite markers for the neotropical tree species Dipteryx alata (Fabaceae). Am J Bot 99:e72–e73. doi: 10.3732/ajb.1100377

Soltis DE, Gitzendanner MA and Soltis PS (2007) A 567-Taxon Data Set for Angiosperms: The Challenges Posed by Bayesian Analyses of Large Data Sets. Int J Plant Sci 168:137–157. doi: 10.1086/509788

Telles MPC, Silva JB, Resende LV, Vianello RP, Chaves LJ, Soares TN and Collevatti RG (2013) Development and characterization of new microsatellites for Eugenia dysenterica DC (Myrtaceae). Genet Mol Res 12:3124–3127. doi: 10.4238/2013.February.6.3

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S (2017) GeSeq - Versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6–W11. doi: 10.1093/nar/gkx391

Tunholi VP, Ramos MA and Scariot A (2013) Availability and use of woody plants in a agrarian reform settlement in the. 27:604–612.

Wyman SK, Jansen RK and Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255. doi: 10.1093/bioinformatics/bth352

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S et al. (2012) Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci 109:17519–17524. doi: 10.1073/pnas.1205818109

Tables:

Table 1. Comparative chloroplast genome structural features in 10 species from Malpighiales order. LSC: Large Single Copy region; SSC: Small Single Copy region; IR: Inverted Repeats regions and bp: base pair.

Figure legends:

Figure 1. Chloroplast genome map of *Caryocar brasiliense*. The genes drawn outside and inside of the circle are transcribed in clockwise and counterclockwise directions, respectively. Genes were colored based on their functional groups. The inner circle shows the quadripartite structure of the chloroplast: small single copy (SSC), large single copy (LSC) and a pair of inverted repeats (IRa and IRb). The gray ring marks the GC content with the inner circle marking a 50% threshold. Genes that have introns were marked with "*" and pseudogenes were marked with "#".

Figure 2. Comparison of the junctions involving Inverted Repeat (IRa and IRb) regions with Large Single Copy (LSC) and Small Single Copy (SSC) regions among ten chloroplast genomes of Malpighiales. The IR regions are represented in yellow whereas LSC and SSC in blue and orange, respectively. The white boxes represent the genes present in each region. The arrows represent the distance (in base pairs) of genes from the junction site between regions.

Figure 3. Phylogenetic tree reconstruction based on 52 taxa using bayesian inference based on 76 protein-coding chloroplast genes. Numbers represent the Bayesian posterior probability given to each node. The bars on the right represents the botanic families of species.

Supplementary material:

Figure S1: Distribution of k-mers in *Caryocar brasiliense* chloroplast genome. Red arrows evidenced the Inverted repeat boundaries and the well sequencing of whole chloroplast genome.

Figure S2. Alignment view of chloroplast genomes of Malpighiales order using *Jatropha curcas* (Euphorbiaceae) as reference. This figure was draw using Mvista software. Grey arrows above the alignment indicates gene orientation. Pink regions represents CNS (Conserved Non-coding Regions). A threshold of 50% identity was used for the plots. Top and bottom of each horizontal bar represents a range of 50% to 100% of identity. Cbr: *Caryocar brasiliense*; Cic: *Chrysobalanus icaco*; Bco: *Byrsonima coccolobifolia*; Mes: *Manihot esculenta* and Ped: *Passiflora edulis*.

Figure S3. Repeat and comparative analysis in *Caryocar brasiliense* chloroplast genome. A) Comparative total number of repeat sequences among Malpighiales chloroplast genome species; B) Comparative repeat types among Malpighiales chloroplast genome species; C) Frequency of microsatellites motifs in *C. brasiliense* chloroplast genome and D) Distribution of microsatellites in *C. brasiliense* chloroplast genome.

Table S1. Gene content and classification of *Caryocar brasiliense* chloroplast genome.

Table S2. Comparative chloroplast genome gene features in 10 species from Malpighiales order. rRNA: ribosomal RNA; tRNA: transfer RNA; CDS: coding sequences.

Table S3: Frequency of types of simple sequence repeats based on its motif length in *Caryocar brasiliense* chloroplast genome.

Table S4: Simple sequence repeats identified in *Caryocar brasiliense* chloroplast genome sequence.

Table S5: Species list used in comparative analysis of Malpighiales order chloroplast genomes.

Table S6: Species list used in phylogenetic analysis of Malpighiales order.

Table 1. Comparative chloroplast genome structural features in 10 species from Malpighiales order. LSC: Large Single Copy region; SSC: Small Single Copy region; IR: Inverted Repeats regions and bp: base pair.

Species	**Family**	**Genome size (bp)**	**LSC (bp)**	**SSC (bp)**	**IR (bp)**	**GC(%)**
Caryocar brasiliense	Caryocaraceae	165,793	84,137	11,852	34,902	36.7
Garcinia mangostana	Clusiaceae	158,179	86,458	17,703	27,009	36.1
Chrysobalanus icaco	Chrysobalanaceae	163,937	89,188	19,817	26,885	36.2
Erythroxylum novogranatense	Erythroxylaceae	163,937	91,384	18,137	27,208	35.9
Manihot esculenta	Euphorbiaceae	161,453	89,295	18,250	26,954	35.9
Linum usitatissimum	Linaceae	156,721	81,767	10,974	31,990	37.5
Byrsonima coccolobifolia	Malpighiaceae	160,329	88,524	17,833	26,986	36.8
Passiflora edulis	Passifloraceae	151,406	85,720	13,378	26,154	37.0
Populus tremula	Salicaceae	156,067	84,367	16,670	27,509	36.8
Viola seoulensis	Violaceae	156,507	85,691	18,008	26,404	36.3

Figures:

Figure 1. Chloroplast genome map of *Caryocar brasiliense*. The genes drawn outside and inside of the circle are transcribed in clockwise and counterclockwise directions, respectively. Genes were colored based on their functional groups. The inner circle shows the quadripartite structure of the chloroplast: small single copy (SSC), large single copy (LSC) and a pair of inverted repeats (IRa and IRb). The gray ring marks the GC content with the inner circle marking a 50% threshold. Genes that have introns were marked with "*" and pseudogenes were marked with "#".

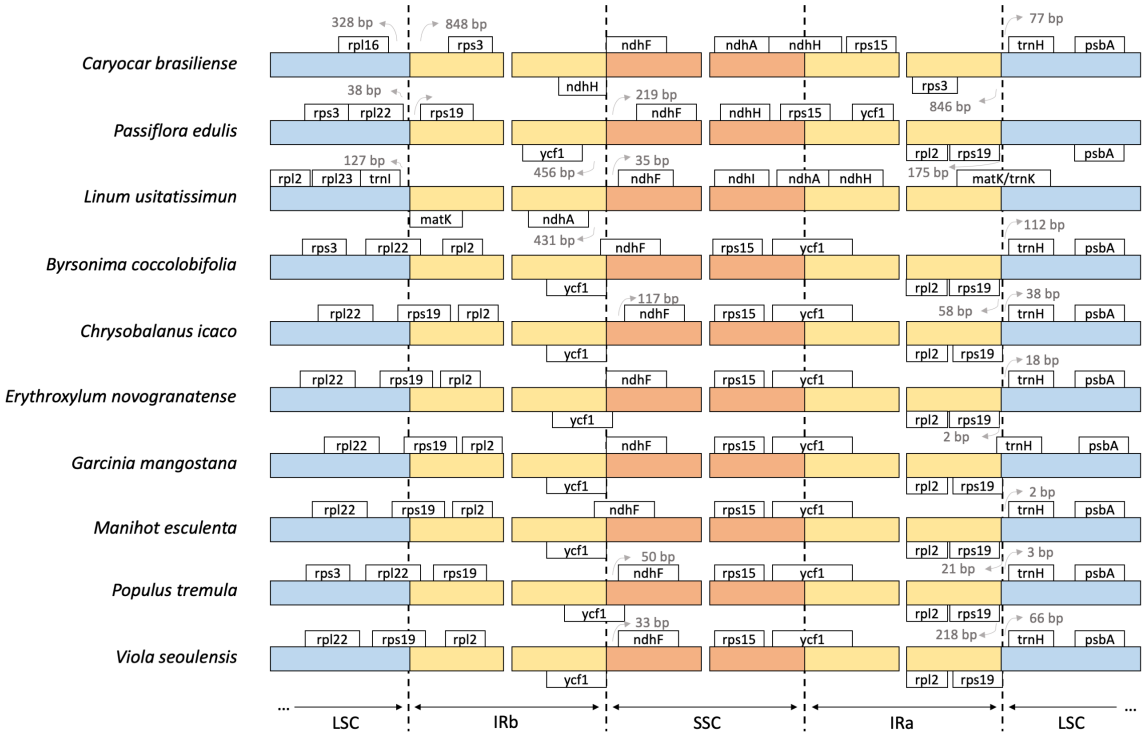


Figure 2. Comparison of the junctions involving Inverted Repeat (IRa and IRb) regions with Large Single Copy (LSC) and Small Single Copy (SSC) regions among ten chloroplast genomes of Malpighiales. The IR regions are represented in yellow whereas LSC and SSC in blue and orange, respectively. The white boxes represent the genes present in each region. The arrows represent the distance (in base pairs) of genes from the junction site between regions.

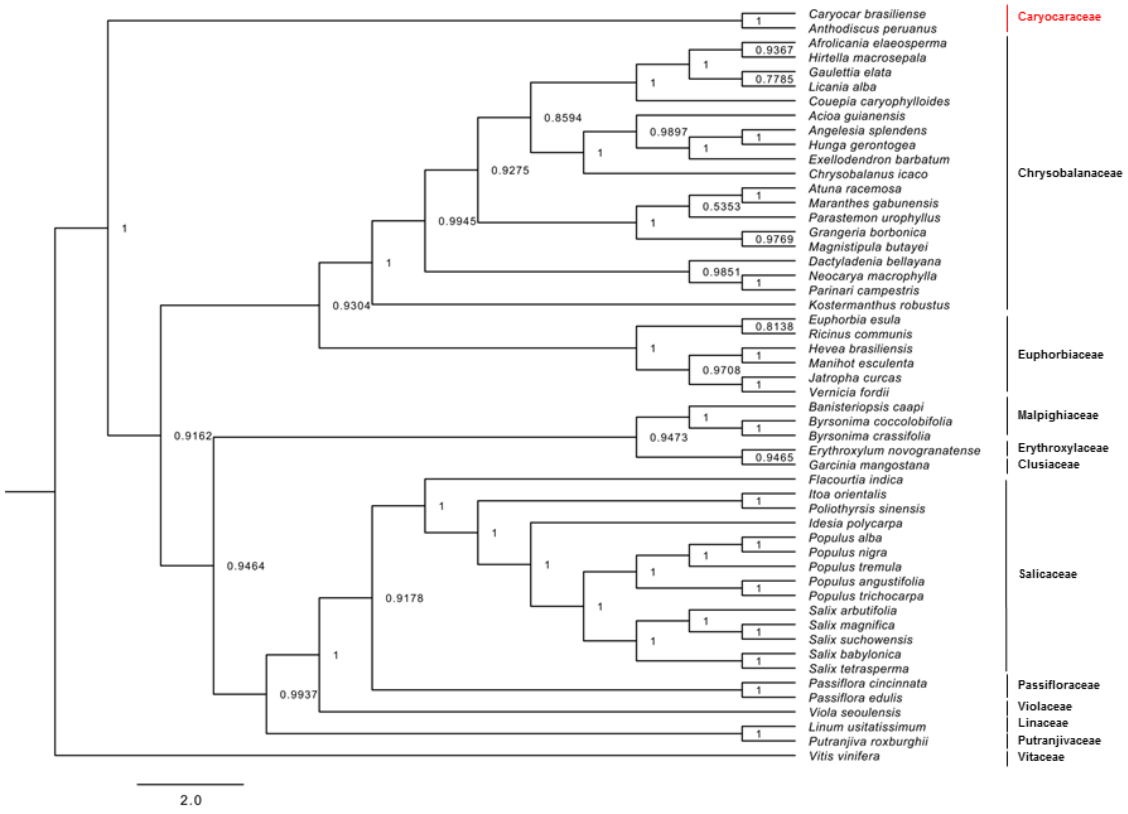


Figure 3. Phylogenetic tree reconstruction based on 52 taxa using bayesian inference based on 76 protein-coding chloroplast genes. Numbers represent the Bayesian posterior probability given to each node. The bars on the right represents the botanic families of species.

Supplement Material:

1 Methods Detailing

1.1 DNA sampling and sequencing

Fresh leaves were collected from a tree at Escola de Agronomia, Universidade Federal de Goiás, Goiânia, Goiás, Brazil. The total DNA was extracted from leaves using CTAB protocol. The quality of DNA was measured by Nanodrop and the quantity by Qbit and agarose gel. The sample was sent to Centro de Genômica Funcional ESALQ-USP core-facility for sequencing. An Illumina paired-end 2x100 bp library was constructed and forwarded for sequencing in an Illumina HiSeq2000 platform.

1.2 Sequencing quality control, assembly and validation

Raw reads were evaluated for base quality sequencing and sequencing adapters presence using FastQC software (Andrews 2010). The quality control was performed using Trimmomatic (Bolger et al. 2014) software with the options ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 and SLIDEWINDOW: 4:30, that is, at least a mean phred-score of 30 every four bases. The high-quality reads were taken for a *de novo* chloroplast genome assembly in NOVOPlasty v.2.7.1 software (Dierckxsens et al. 2017). The *Caryocar brasiliense* psbA gene sequence (GenBank: EU350266.1) was used as seed to extend the chloroplast genome sequence in assembly analysis. The assembled chloroplast genome and the junctions between inverted repeats and single copy regions were validated in a k-mer coverage analysis using Jellyfish software (Marçais and Kingsford 2011) (Figure S1).

1.3 Chloroplast structure and gene annotation

We performed a gene annotation of *Caryocar brasiliense* chloroplast genome using CHLOROBOX GeSeq (Tillich et al. 2017) and DOGMA - Dual Organellar GenoMe Annotator (Wyman et al. 2004) software. Predicted genes were curated using BLAT and BLAST searches in a database with chloroplastidial Embryophyta CDS and RNA reference sequences. ARAGORN v1.2.38 (Laslett and Canback 2004) and tRNAscan-SE v2.0 (Lowe and Eddy 1997; Lowe and Chan 2016) software were also used to predict tRNA sequences. ARAGORN was configured in "Bacterial/Plant plastid" genetic code mode and with a maximum intron length of 3000 bp. tRNAscan was configured in "Organellar tRNAs" and a cut-off score for reporting tRNAs of 15. The circular chloroplast genome map was created using OrganellarGenomeDRAW (Lohse et al. 2013).

1.4 Comparative analysis among Malpighiales species

Comparative analysis of structure and composition of chloroplast genome sequences were performed using a dataset compounded by nine species in Malpighiales order: *Byrsonima coccolobifolia* (NC_037191.1), *Chrysobalanus icaco* (NC_024061.1), *Erythroxylum novogranatense* (NC_030601.1), *Garcinia mangostana* (NC_036341.1), *Linum usitatissimum* (NC_036356.1), *Manihot esculenta* (NC_010433.1), *Passiflora edulis* (NC_034285.1), *Populus tremula* (NC_027425.1) and *Viola seoulensis* (NC_026986.1). The species dataset represent all families in Malpighiales with chloroplast genomes sequenced until now. The plastid genomes were retrieved from Genbank, as well as its genome annotation information. The dataset was used in whole-genome alignment view, repeat sequence and IR boundaries analysis. Comparisons were performed using Geneious software (Duran et al. 2012).

An alignment view of chloroplast genomes of Malpighiales were drawn using mVISTA software (Frazer et al. 2004). *Jatropha curcas* (Euphorbiaceae) (NC_012224.1) chloroplast genome was used as reference (Asif et al. 2010). The analysis was conducted in Shuffle-LAGAN mode (Global pair-wise alignment of finished sequences) and probability threshold of 0.5. To facilitate the visualization, a resumed alignment figure was made using only five species (Cbr: *Caryocar brasiliense*; Cic: *Chrysobalanus icaco*; Bco: *Byrsonima coccolobifolia*; Mes: *Manihot esculenta*; Ped: *Passiflora edulis*).

1.5 Repeat sequence analysis

Simple sequence repeats (SSR) or microsatellite regions were predicted in *C. brasiliense* chloroplast genome using IMEx - Imperfect Microsatellite Extractor (Mudunuri and Nagarajaram 2007). We used the following minimum repeat number criteria: ten units for mononucleotide, five units for dinucleotide, four units for trinucleotide and three units for tetra, penta and hexanucleotides, respectively. All repeat sequence positions were retrieved using Geneious software (Duran et al. 2012).

We looked for repeat sequence elements in *Caryocar brasiliense* chloroplast genome using REPuter software (Kurtz 2002). Forward, reverse, complement and palindromic repeats types were searched using a minimal repeat size of 30 bp and a Hamming distance of 3 (so, sequence identities ≥ 90%). We used the same search criteria to identify repeat sequence elements in the chloroplast genomes of *Byrsonima coccolobifolia, Chrysobalanus icaco*, *Erythroxylum novogranatense, Garcinia mangostana, Linum usitatissimum, Manihot esculenta, Passiflora edulis, Populus tremula* and *Viola seoulensis*. We used this information in a quantitative comparative analysis among families from Malpighiales order.

1.6 Phylogenetic analyses

We performed a bayesian phylogenetic analysis among some Malpighiales species using 76 protein-coding gene sequences. We retrieved coding sequences (CDS) from 49 complete chloroplast genome sequences from Nucleotide NCBI (National Center for Biotechnology Information) database representing all the families in Malpighiales order that had their chloroplast genomes sequence until now: Chrysobalanaceae, Clusiaceae, Erythroxylaceae, Euphorbiaceae, Linaceae, Malpighiaceae, Passifloraceae, Salicaceae e Violaceae. Additionally, we also collect chloroplast gene sequences from *Anthodiscus peruanus* (Caryocaraceae) and *Putranjiva roxburghii* (Putranjivaceae) (Xi et al. 2012). *Vitis vinifera* L. (Vitaceae) chloroplast gene sequences were used as outgroup. The GenBank accessions are listed in supplement material (Table S6).

The 76 shared protein-coding gene were aligned separately by gene using MAFFT v. 7 (Katoh and Standley 2013) server software. Alignment files were concatenated in a single matrix using Sequence Matrix software (Vaidya et al. 2011). Informative site were extracted from matrix using GBlocks v. 0.91b (Cruickshank 2000) and was used as input in jModelTest v. 2.1.10 (Posada 2008; Darriba et al. 2012) to test the best-fitting evolution model for phylogeny estimation based on the Akaike Information Criterion (AIC). The model GTR+G+I was chose like evolutionary model in phylogenetic analysis using a Bayesian inference approach in MrBayes v. 3.2 software (Huelsenbeck and Ronquist 2001; Darling et al. 2012). This analysis was conducted using two independent Markov chain Monte Carlo (MCMC) chains with three hot chains and one cold chain each and 2 x 5,000,000 generations. Phylogenetic trees were sampled every 1,000 generations and the first 25% were discarded as burn-in. The remaining trees were used to construct majority-rule consensus. The MCMC convergence was assumed when the average standard deviation

of split frequencies reached 0.01 or less. The runs were evaluated using Tracer software (Rambaut et al. 2018). Phylogenetic tree visualization was drawn using FigTree (Rambaut A, 2006).

References of Methods Detailing

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Asif MH, Mantri SS, Sharma A, Srivastava A, Trivedi I, Gupta P, Mohanty CS, Sawant S V. and Tuli R (2010) Complete sequence and organisation of the Jatropha curcas (Euphorbiaceae) chloroplast genome. Tree Genet Genomes 6:941–952. doi: 10.1007/s11295-010-0303-0

Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi: 10.1093/bioinformatics/btu170

Cruickshank R (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–52.

Darling A, Ronquist F, Ayres DL, Larget B, Liu L, Teslenko M, Suchard MA, Huelsenbeck JP, Höhna S and van der Mark P (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol 61:539–542. doi: 10.1093/sysbio/sys029

Darriba D, Taboada GL, Doallo R and Posada D (2012) JModelTest 2: More models, new heuristics and parallel computing. Nat Methods 9:772. doi: 10.1038/nmeth.2109

Dierckxsens N, Mardulyn P and Smits G (2017) NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. doi: 10.1093/nar/gkw955

Duran C, Markowitz S, Moir R, Cooper A, Ashton B, Drummond A, Buxton S, Sturrock S, Wilson A, Thierer T et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649. doi: 10.1093/bioinformatics/bts199

Frazer KA, Pachter L, Poliakov A, Rubin EM and Dubchak I (2004) VISTA: Computational tools for comparative genomics. Nucleic Acids Res 32:273–279. doi: 10.1093/nar/gkh458

Huelsenbeck JP and Ronquist F (2001) MrBAYES; Bayesian inference for phylogeny. Bioinformatics 17:754–755. doi: 10.1093/bioinformatics/17.8.754

Katoh K and Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780. doi: 10.1093/molbev/mst010

Kurtz S (2002) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642. doi: 10.1093/nar/29.22.4633

Laslett D and Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16. doi: 10.1093/nar/gkh152

Lohse M, Drechsel O, Kahlau S and Bock R (2013) OrganellarGenomeDRAW--a suite of tools for

generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41:575–581. doi: 10.1093/nar/gkt289

Lowe TM and Chan PP (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44:W54–W57. doi: 10.1093/nar/gkw413

Lowe TM and Eddy SR (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Res 25:0955–0964. doi: 10.1093/nar/25.5.0955

Marçais G and Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–770. doi: 10.1093/bioinformatics/btr011

Mudunuri SB and Nagarajaram HA (2007) IMEx: Imperfect microsatellite extractor. Bioinformatics 23:1181–1187. doi: 10.1093/bioinformatics/btm097

Posada D (2008) jModelTest: Phylogenetic model averaging. Mol Biol Evol 25:1253–1256. doi: 10.1093/molbev/msn083

Rambaut A, Drummond AJ, Xie D, Baele G and Suchard MA (2018) Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol 67:901–904. doi: 10.1093/sysbio/syy032

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S (2017) GeSeq - Versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6–W11. doi: 10.1093/nar/gkx391

Vaidya G, Lohman DJ and Meier R (2011) SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27:171–180. doi: 10.1111/j.1096-0031.2010.00329.x

Wyman SK, Jansen RK and Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255. doi: 10.1093/bioinformatics/bth352

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S et al. (2012) Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci 109:17519–17524. doi: 10.1073/pnas.1205818109

Table S1. Gene content and classification of Caryocar brasiliense chloroplast genome.

Gene category	Gene group	Gene name
Photosynthesis	Photosystem I	psaA, psaB, psaC, psaI, psaJ
	Photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbT, psbZ, pbf1
	Cytochrome b6/f complex	petA, petB, petD, petG, petL, petN
	ATP synthase	atpA, atpB, atpE, atpF[a], atpH, atpI
	Cytochrome c synthesis	ccsA
	Assembly/stability of photosystem I	ycf3[a], ycf4
	NADPH dehydrogenase	ndhA[a], ndhB[a,b], ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
	Rubisco	rbcL
Transcription and translation	Transcription	rpoA, rpoB, rpoC1[a], rpoC2
	Ribosomal proteins	rpl2[a,b], rpl14, rpl16, rpl20, rpl22[b], rpl23[b], rpl33, rpl36, rps2, rps3[b], rps4, rps7[b], rps8, rps11, rps12[a], rps14, rps15[b], rps18, rps19[b]
Non-coding RNA genes	Ribosomal RNA	rrn4.5[b], rrn5[b], rrn16[b], rrn23[b]
	Transfer RNA	TrnA-UGC[a,b], trnC-GCA, trnD-GUC, trnE-UUC[a,b], trnF-GAA, trnfM-CAU, trnG-UCC, trnH-GUG, trnI-CAU, trnI-GAU[a,b], trnK-UUU, trnL-CAA[b], trnL-UAA, trnL-UAG, trnM-CAU[b], trnN-GUU, trnP-GGG, trnP-UGG, trnQ-UUG, trnR-ACG[b], trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC[b], trnV-UAC, trnW-CCA, trnY-GUA
Other genes	RNA processing	matK
	Carbon metabolism	cemA
	Fatty acid synthesis	accD
	Proteolysis	clpP[a]
	Translation initiation factor	InfA
	Component of TIC complex	ycf1[b]
Genes of unknown function	Conserved reading frames	ycf2[b]
Pseudogenes	-	ndhH, psaA, psaB, psbA

[a] Gene containing intron; [b] Gene located in inverted repeat regions;

Table S2. Comparative chloroplast genome gene features in 10 species from Malpighiales order. rRNA: ribosomal RNA; tRNA: transfer RNA; CDS: coding sequences.

Species	Family	Number of genes	rRNA	tRNA	CDS	Pseudogenes
Caryocar brasiliense	Caryocaraceae	136	8	37	87	4
Garcinia mangostana	Clusiaceae	130	8	37	83	2
Chrysobalanus icaco	Chrysobalanaceae	130	8	37	83	2
Erythroxylum novogranatense	Erythroxylaceae	131	4	38	85	4
Manihot esculenta	Euphorbiaceae	131	4	38	83	6
Linum usitatissimum	Linaceae	133	8	37	83	5
Byrsonima coccolobifolia	Malpighiaceae	139	8	37	88	6
Passiflora edulis	Passifloraceae	133	8	36	76	13
Populus tremula	Salicaceae	123	8	29	85	1
Viola seoulensis	Violaceae	131	8	37	84	2

Table S3: Frequency of types of simple sequence repeats based on its motif length in *Caryocar brasiliense* chloroplast genome.

Repeat unit	frequency	%
mono	52	60,47
di	11	12,79
tri	5	5,81
tretra	12	13,95
penta	4	4,65
hexa	2	2,33
Total	86	100,00

Table S4: Simple sequence repeats identified in *Caryocar brasiliense* chloroplast genome sequence.

Number	Consensus	Rep. Size	Iterations	Tract-size	Start	End	Region
1	TTGAA	5	3	15	833	847	LSC
2	A	1	13	13	1389	1401	LSC
3	A	1	11	11	1901	1911	LSC
4	AGAA	4	3	12	7270	7281	LSC

5	TTTA	4	3	12	10049	10060	LSC
6	A	1	14	14	10533	10546	LSC
7	A	1	10	10	10647	10656	LSC
8	T	1	11	11	10996	11006	LSC
9	A	1	10	10	11648	11657	LSC
10	A	1	10	10	11689	11698	LSC
11	AT	2	5	10	14406	14415	LSC
12	T	1	10	10	17884	17893	LSC
13	A	1	10	10	21267	21276	LSC
14	TAT	3	4	12	22775	22786	LSC
15	TAA	3	5	15	22859	22873	LSC
16	CAAA	4	3	12	31910	31921	LSC
17	A	1	10	10	32444	32453	LSC
18	T	1	10	10	35206	35215	LSC
19	AT	2	5	10	35262	35271	LSC
20	AT	2	6	12	37507	37518	LSC
21	AT	2	5	10	37580	37589	LSC
22	T	1	10	10	38485	38494	LSC
23	TA	2	5	10	39088	39097	LSC
24	T	1	10	10	41181	41190	LSC
25	A	1	12	12	41192	41203	LSC
26	T	1	10	10	41908	41917	LSC
27	TA	2	5	10	43796	43805	LSC
28	T	1	13	13	48386	48398	LSC
29	AT	2	5	10	49768	49777	LSC
30	A	1	10	10	52433	52442	LSC
31	A	1	10	10	52816	52825	LSC
32	T	1	10	10	56118	56127	LSC
33	A	1	10	10	57615	57624	LSC
34	C	1	13	13	57638	57650	LSC

35	A	1	10	10	60015	60024	LSC
36	T	1	10	10	60583	60592	LSC
37	ACAA	4	3	12	60676	60687	LSC
38	T	1	10	10	60921	60930	LSC
39	ATA	3	4	12	68680	68691	LSC
40	TCTT	4	3	12	75435	75446	LSC
41	A	1	10	10	76659	76668	LSC
42	A	1	11	11	76808	76818	LSC
43	T	1	10	10	77087	77096	LSC
44	T	1	12	12	79216	79227	LSC
45	TCAAT	5	3	15	79258	79272	LSC
46	AGAA	4	3	12	79945	79956	LSC
47	A	1	11	11	81636	81646	LSC
48	T	1	10	10	83850	83859	LSC
49	T	1	16	16	86456	86471	IR
50	AT	2	5	10	96737	96746	IR
51	T	1	12	12	101288	101299	IR
52	TATT	4	3	12	101348	101359	IR
53	AATGGA	6	3	18	102377	102394	IR
54	A	1	11	11	110421	110431	IR
55	AG	2	5	10	110512	110521	IR
56	A	1	10	10	114492	114501	IR
57	A	1	10	10	114712	114721	IR
58	A	1	10	10	115777	115786	IR
59	A	1	10	10	116260	116269	IR
60	T	1	10	10	117616	117625	IR
61	TAT	3	4	12	117873	117884	SSC
62	TTTTA	5	3	15	120631	120645	SSC
63	TCTT	4	3	12	120732	120743	SSC
64	TGGT	4	3	12	122433	122444	SSC

65	A	1	11	11	125343	125353	SSC
66	TATT	4	3	12	125730	125741	SSC
67	A	1	10	10	125954	125963	SSC
68	TTAA	4	3	12	127452	127463	SSC
69	A	1	11	11	127629	127639	SSC
70	A	1	10	10	128177	128186	SSC
71	T	1	10	10	128219	128228	SSC
72	T	1	10	10	128374	128383	SSC
73	AAT	3	4	12	132046	132057	IR
74	A	1	10	10	132306	132315	IR
75	T	1	10	10	133662	133671	IR
76	T	1	10	10	134145	134154	IR
77	T	1	10	10	135210	135219	IR
78	T	1	10	10	135430	135439	IR
79	CT	2	5	10	139410	139419	IR
80	T	1	11	11	139500	139510	IR
81	TCCATT	6	3	18	147537	147554	IR
82	ATAA	4	3	12	148573	148584	IR
83	A	1	12	12	148632	148643	IR
84	AT	2	5	10	153185	153194	IR
85	A	1	16	16	163460	163475	IR

Table S5: Species list used in comparative analysis of Malpighiales order chloroplast genomes.

Species	Family	Reference	NCBI
Caryocar brasiliense	Caryocaraceae	This work	-
Garcinia mangostana	Clusiaceae	Jo et al., 2017	NC_036341.1
Chrysobalanus icaco	Chrysobalanaceae	Malé et al., 2014	NC_024061.1
Erythroxylum novogranatense	Erythroxylaceae	Unpublished	NC_030601.1
Manihot esculenta	Euphorbiaceae	Daniell et al., 2008	NC_010433.1
Linum usitatissimum	Linaceae	Lopes et al., 2017	NC_036356.1
Byrsonima coccolobifolia	Malpighiaceae	Menezes et al., 2018	NC_037191.1
Passiflora edulis	Passifloraceae	Cauz-Santos et al., 2017	NC_034285.1
Populus tremula	Salicaceae	Kersten et pal., 2016	NC_027425.1
Viola seoulensis	Violaceae	Cheon et al., 2015	NC_026986.1

Table S6: Species list used in phylogenetic analysis of Malpighiales order.

ID	**Family**	**Species**	**Code**	**Reference**
1	Caryocaraceae	*Caryocar brasiliense*	Cbr	This work
2	Caryocaraceae	*Anthodiscus peruanus*	Ape	Xi et al., 2012
3	Clusiaceae	*Garcinia mangostana*	Gma	NC_036341.1
4	Chrysobalanaceae	*Acioa guianensis*	Agu	NC_030534.1
5	Chrysobalanaceae	*Afrolicania elaeosperma*	Ael	NC_030544.1
6	Chrysobalanaceae	*Angelesia splendens*	Asp	NC_030545.1
7	Chrysobalanaceae	*Atuna racemosa*	Ara	NC_030546.1
8	Chrysobalanaceae	*Chrysobalanus icaco*	Cic	NC_024061.1
9	Chrysobalanaceae	*Couepia caryophylloides*	Cca	NC_030547.1
10	Chrysobalanaceae	*Dactyladenia bellayana*	Dbe	NC_030555.1
11	Chrysobalanaceae	*Exellodendron barbatum*	Eba	NC_030558.1
12	Chrysobalanaceae	*Gaulettia elata*	Gel	NC_030559.1
13	Chrysobalanaceae	*Grangeria borbonica*	Gbo	NC_030560.1

14	Chrysobalanaceae	*Hirtella macrosepala*	Hma	NC_030561.1
15	Chrysobalanaceae	*Hunga gerontogea*	Hge	NC_030564.1
16	Chrysobalanaceae	*Kostermanthus robustus*	Kro	NC_030565.1
17	Chrysobalanaceae	*Licania alba*	Lal	NC_024064.1
18	Chrysobalanaceae	*Magnistipula butayei*	Mbu	NC_030576.1
19	Chrysobalanaceae	*Maranthes gabunensis*	Mga	NC_030577.1
20	Chrysobalanaceae	*Neocarya macrophylla*	Nma	NC_030580.1
21	Chrysobalanaceae	*Parastemon urophyllus*	Pur	NC_030517.1
22	Chrysobalanaceae	*Parinari campestris*	Pca	NC_024067.1
23	Erythroxylaceae	*Erythroxylum novogranatense*	Eno	NC_030601.1
24	Euphorbiaceae	*Euphorbia esula*	Ees	NC_033910.1
25	Euphorbiaceae	*Hevea brasiliensis*	Hbr	NC_015308.1
26	Euphorbiaceae	*Jatropha curcas*	Jcu	NC_012224.1
27	Euphorbiaceae	*Manihot esculenta*	Mes	NC_010433.1
28	Euphorbiaceae	*Ricinus communis*	Rco	NC_016736.1
29	Euphorbiaceae	*Vernicia fordii*	Vfo	NC_034803.1
30	Linaceae	*Linum usitatissimum*	Lus	NC_036356.1
31	Malpighiaceae	*Banisteriopsis caapi*	Bca	NC_037945.1
32	Malpighiaceae	*Byrsonima coccolobifolia*	Bco	NC_037191.1
33	Malpighiaceae	*Byrsonima crassifolia*	Bcr	NC_037192.1
34	Passifloraceae	*Passiflora cincinnata*	Pci	NC_037690.1
35	Passifloraceae	*Passiflora edulis*	Ped	NC_034285.1
36	Salicaceae	*Flacourtia indica*	Fin	NC_037410.1
37	Salicaceae	*Idesia polycarpa*	Ipo	NC_032060.1
38	Salicaceae	*Itoa orientalis*	Ior	NC_037411.1
39	Salicaceae	*Poliothyrsis sinensis*	Psi	NC_037412.1
40	Salicaceae	*Populus alba*	Pal	NC_008235.1
41	Salicaceae	*Populus angustifolia*	Pan	NC_037413.1
42	Salicaceae	*Populus tremula*	Pte	NC_027425.1
43	Salicaceae	*Populus nigra*	Pni	NC_037416.1
44	Salicaceae	*Populus trichocarpa*	Ptr	NC_009143.1
45	Salicaceae	*Salix arbutifolia*	Sar	NC_036718.1
46	Salicaceae	*Salix babylonica*	Sba	NC_028350.1
47	Salicaceae	*Salix magnifica*	Sma	NC_037424.1
48	Salicaceae	*Salix suchowensis*	Ssu	NC_026462.1
49	Salicaceae	*Salix tetrasperma*	Ste	NC_035744.1
50	Violaceae	*Viola seoulensis*	Vse	NC_026986.1
51	Vitaceae	*Vitis vinifera*	Vvi	NC_007957.1
52	Putranjivaceae	*Putranjiva roxburghii*	Pro	Xi et al., 2012

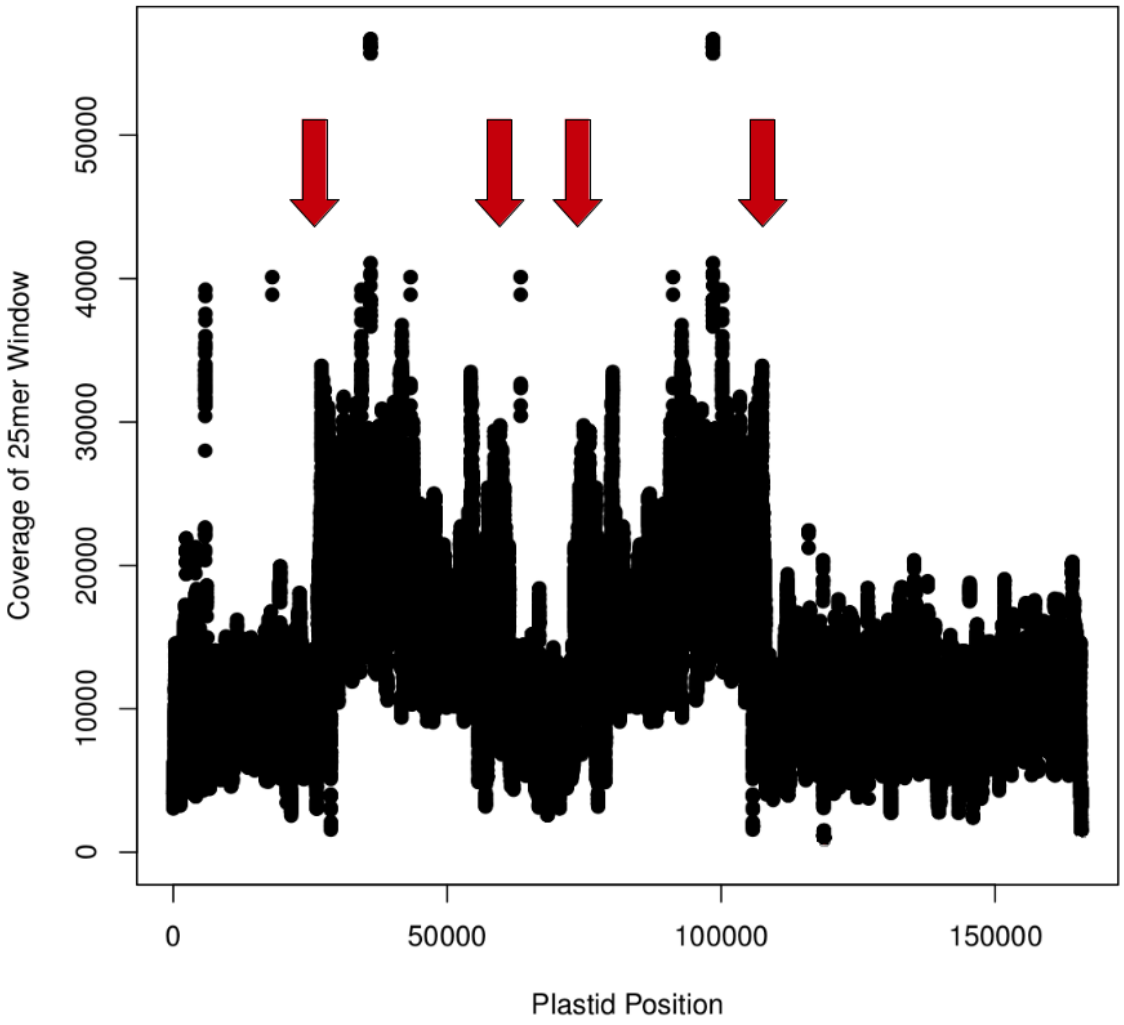


Figure S2: Distribution of k-mers in Caryocar brasiliense chloroplast genome. Red arrows evidenced the Inverted repeat boundaries and the well sequencing of whole chloroplast genome.

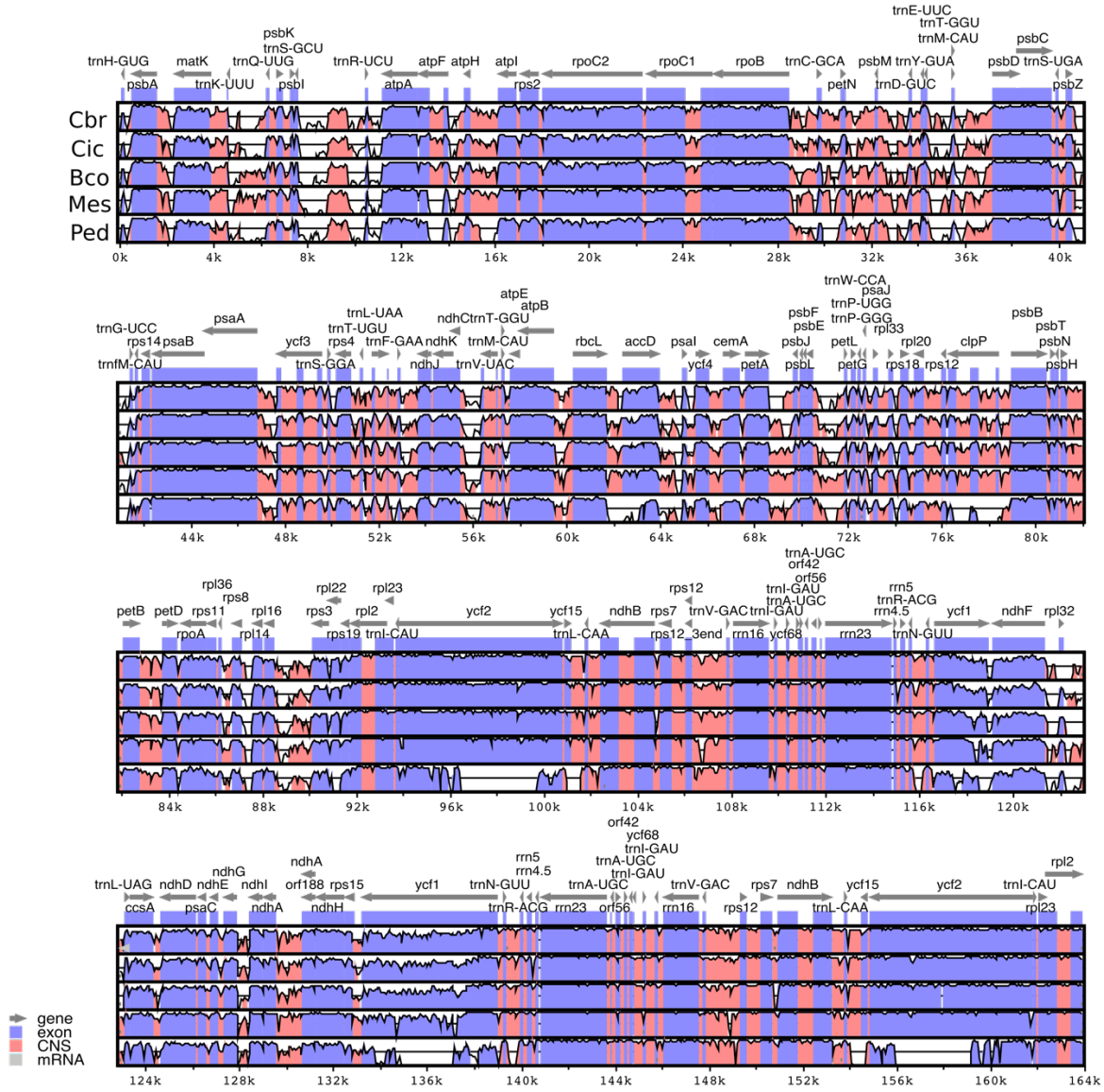


Figure S2. Alignment view of chloroplast genomes of Malpighiales order using *Jatropha curcas* (Euphorbiaceae) as reference. This figure was draw using Mvista software. Grey arrows above the alignment indicates gene orientation. Pink regions represents CNS (Conserved Non-coding Regions). A threshold of 50% identity was used for the plots. Top and bottom of each horizontal bar represents a range of 50% to 100% of identity. Cbr: *Caryocar brasiliense*; Cic: *Chrysobalanus icaco*; Bco: *Byrsonima coccolobifolia*; Mes: *Manihot esculenta* and Ped: *Passiflora edulis*.

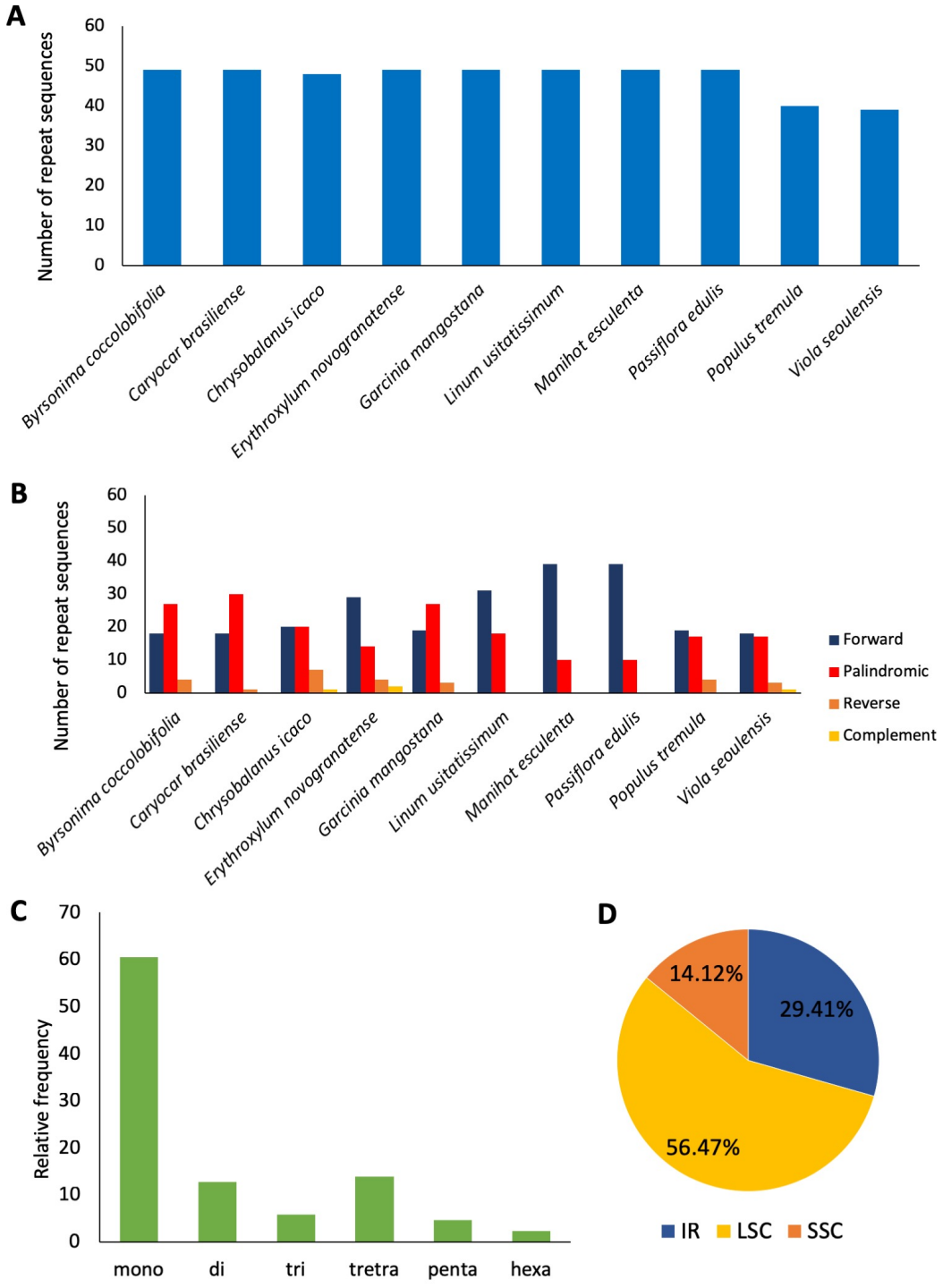


Figure S3. Repeat and comparative analysis in *Caryocar brasiliense* chloroplast genome. A) Comparative total number of repeat sequences among Malpighiales chloroplast genome species; B) Comparative repeat types among Malpighiales chloroplast genome species; C) Frequency of microsatellites motifs in *C. brasiliense* chloroplast genome and D) Distribution of microsatellites in *C. brasiliense* chloroplast genome.

5 CONCLUSÃO GERAL

- Uma visão geral sobre a produção de conhecimento científico para a família Caryocaraceae foi realizado e mostra uma heterogeneidade quanto ao conhecimento produzido até o momento para as espécies pertencentes a esse grupo;
- Nesse trabalho foram gerados e disponibilizados em banco de dados público os primeiros recursos genômicos em larga escala para o Pequizeiro (*Caryocar brasiliense* Camb.);
- Os iniciadores desenhados para amplificação, em multiplex, de regiões microssatélites podem testados e padronizados para disponibilizar novos marcadores para serem utilizados em estudos de genética de populações de *C. brasiliense*;
- O genoma do cloroplasto de *C. brasiliense*, o primeiro da família Caryocaraceae, se configura como um importante recurso genômico permitindo uma melhor compreensão da evolução de espécies da ordem Malpighiales.

6 REFERÊNCIAS GERAIS

ANGELONI, F. et al. De novo transcriptome characterization and development of genomic tools for Scabiosa columbaria L. using next-generation sequencing techniques. p. 662–674, 2011.

CHENG, S. et al. 10KP: A phylodiverse genome sequencing plan. **GigaScience**, v. 7, n. 3, p. 1–9, 2018.

COLLINS, F. S. The Human Genome Project: Lessons from Large-Scale Biology. **Science**, v. 300, n. 5617, p. 286–290, 11 abr. 2003.

COLOMBO, N. B. R. et al. Caryocar brasiliense camb protects against genomic and oxidative damage in urethane-induced lung carcinogenesis. **Brazilian Journal of Medical and Biological Research**, v. 48, n. 9, p. 852–862, 2015.

DAWSON, I. K. et al. Does biotechnology have a role in the promotion of underutilised crops? **Food Policy**, v. 34, n. 4, p. 319–328, ago. 2009.

DE ARAUJO, F. D. A review of caryocar brasiliense (caryocaraceae)—an economically valuable species of the central brazilian cerrados. **Economic Botany**, v. 49, n. 1, p. 40–48, jan. 1995.

DE CARVALHO, L. M. et al. Bioinformatics applied to biotechnology: A review towards bioenergy research. **Biomass and Bioenergy**, v. 123, n. March, p. 195–224, 2019.

GARNER, B. A. et al. Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application. **Trends in Ecology and Evolution**, v. 31, n. 2, p. 81–82, 2016.

GOFF, S. A. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). **Science**, v. 296, n. 5565, p. 92–100, 5 abr. 2002.

GOMEZ-CABRERO, D. et al. Data integration in the era of omics : current and future challenges. v. 8, n. Suppl 2, p. 1–10, 2014.

JOSEPH, B.; NAIR, V. M. Woman Innovator in Bioinformatics: Dr. Margaret Oakley Dayhoff. v. 12, n. 01, p. 32–34, 2012.

KAUL, S. et al. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. **Nature**, v. 408, n. 6814, p. 796–815, dez. 2000.

LEITCH, I.; BOTANIC, R.; MARY, L. Q. Plant genomes – progress and prospects. **Royal Botanic Gardens, Kew**, n. May, p. 16–21, 2017.

LEWIN, H. A. et al. Earth BioGenome Project: Sequencing life for the future of life. **Proceedings of the National Academy of Sciences**, v. 115, n. 17, p. 4325–4333, 24 abr. 2018.

LIU, B.; ZHANG, L.; WANG, X. Scientometric profile of global rice research during 1985-2014. **Current Science**, v. 112, n. 5, p. 1003–1011, 2017.

MANEL, S. et al. Genomic resources and their influence on the detection of the signal of positive selection in genome scans. **Molecular Ecology**, v. 25, n. 1, p. 170–184, 2016.

MARIANO, R. G. DE B.; COURI, S.; FREITAS, S. P. Enzymatic technology to improve oil extraction from Caryocar brasiliense camb. (Pequi) Pulp. **Revista Brasileira de Fruticultura**, v. 31, n. 3, p. 637–643, set. 2009.

MAYES, S. et al. The potential for underutilized crops to improve security of food production. **Journal of Experimental Botany**, v. 63, n. 3, p. 1075–1079, 2012.

METZKER, M. L. Sequencing technologies - the next generation. **Nature reviews. Genetics**, v. 11, n. 1, p. 31–46, 2010.

METZKER, M. L. M. L. L. Emerging technologies in DNA sequencing. **Genome Res.**, v. 15, n. 12, p. 1767–76, 2005.

MICHAEL, T. P.; JACKSON, S. The First 50 Plant Genomes. 2013.

MIRANDA-VILELA, A. L. et al. Pequi fruit (Caryocar brasiliense Camb.) pulp oil reduces exercise-induced inflammatory markers and blood pressure of male and female runners. **Nutrition Research**, v. 29, n. 12, p. 850–858, dez. 2009.

OUZOUNIS, C. A. Rise and Demise of Bioinformatics? Promise and Progress. v. 8, n. 4, 2012.

PIANOVSKI, A. R. et al. Uso do óleo de pequi (Caryocar brasiliense) em emulsões

cosméticas: desenvolvimento e avaliação da estabilidade física. **Revista Brasileira de Ciências Farmacêuticas**, v. 44, n. 2, p. 249–259, 2008.

REISS, T. Drug discovery of the future: the implications of the human genome project. **Trends in Biotechnology**, v. 19, n. 12, p. 496–499, 1 dez. 2001.

ROLL, M. M. et al. The pequi pulp oil (Caryocar Brasiliense Camb.) provides protection against aging-related anemia, inflammation and oxidative stress in Swiss mice, especially in females. **Genetics and Molecular Biology**, v. 41, n. 4, p. 858–869, 2018.

SANGER, F. et al. Nucleotide sequence of bacteriophage φX174 DNA. **Nature**, v. 265, n. 5596, p. 687–695, fev. 1977.

SHENDURE, J. et al. DNA sequencing at 40: past, present and future. **Nature**, v. 550, n. 7676, p. 345–353, 11 out. 2017.

TANAKA, K. et al. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. **Nature**, v. 428, p. 653–7, 2004.

TUNHOLI, V. P.; RAMOS, M. A.; SCARIOT, A. Availability and use of woody plants in a agrarian reform settlement in the. v. 27, n. 3, p. 604–612, 2013.

TÜRKTAŞ, M.; KURTOĞLU, K. Y.; DORADO, G. Sequencing of plant genomes – a review. p. 361–376, 2015.

TUSKAN, G. A. et al. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). **Science**, v. 313, n. 5793, p. 1596–1604, 15 set. 2006.

VIEIRA, R. F.; CAMILLO, J.; CORADIN, L. Espécies nativas da flora brasileira de valor econômico atual ou potencial: plantas para o futuro: Região Centro-Oeste. **Embrapa Recursos Genéticos e Biotecnologia-Livro científico (ALICE)**, 2018.

WARD, D. C.; WHITE, D. C. The new “omics” era. Editorial overview. **Curr.Opin.Biotechnol.**, v. 13, p. 11–13, 2002.

YU, J. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). **Science**, v. 296, n. 5565, p. 79–92, 5 abr. 2002.

APÊNDICES

Apêndice A: Artigo de ensino/divulgação científica publicado na revista Genética na Escola sobre polimorfismos e domesticação em plantas;

Apêndice B: Artigo de ensino/divulgação científica publicado na revista Genética na Escola resenhando um livro biográfico do pesquisador James Watson.

O gene *qSH1* e a domesticação do arroz

Rhewter Nunes[1,2], Stela Barros Ribeiro[1,2], Ivone de Bem Oliveira[1,2], Isabela Pavanelli de Souza[1,2], Mariana Pires de Campos Telles[2,3], Alexandre Siqueira Guedes Coelho[1,2]

[1] Laboratório de Genética e Genômica de Plantas, Escola de Agronomia, Universidade Federal de Goiás (UFG), Goiânia, GO, Brasil.
[2] Programa de Pós-Graduação em Genética e Melhoramento de Plantas, Escola de Agronomia, Universidade Federal de Goiás (UFG), Goiânia, GO, Brasil.
[3] Escola de Ciências Agrárias e Biológicas, Pontifícia Universidade Católica de Goiás (PUC-GO) e Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás (UFG), Goiânia, GO, Brasil.

Autor para correspondência: tellesmpc@gmail.com

Diversos fatores genéticos estão envolvidos na variação e, consequentemente, estão relacionados com o processo de domesticação que deu origem a muitas plantas utilizadas atualmente na agricultura. Essas variações podem estar associadas a genes específicos envolvidos no processo de domesticação. Neste artigo apresentamos o gene *qSH1* (*seed shattering*), que está envolvido no controle genético do degrane das sementes do arroz possibilitando discutir como uma pequena variação pode estar relacionada com o processo de domesticação de uma espécie vegetal e, além disso, também discorrer sobre a estrutura e o papel desse gene no arroz e importância do mesmo para as populações humanas.

COMO SURGIU A RELAÇÃO ENTRE HOMENS E PLANTAS?

Um dos eventos mais importantes que levou o homem pré-histórico a conseguir se estabelecer, dando origem a uma sociedade complexa, envolvendo grande número de pessoas, foi a mudança em seu hábito de adquirir recursos para a sobrevivência. Há mais de 10 mil anos, as populações humanas dependiam de recursos provenientes da caça de animais selvagens e da coleta de frutos, raízes e folhas das plantas que ocupavam os ambientes em que as populações se encontravam. Por possuírem esse hábito, foram considerados caçadores-coletores e, provavelmente, apresentavam comportamento nômade para encontrar uma nova área com recursos a cada vez que os recursos da área ocupada naquele momento se exauriam.

Vivendo dessa forma, o tamanho das populações humanas era limitado pela restrição de recursos existentes nas regiões onde elas se encontravam. Nessas condições, a seleção de alimentos melhores devia ser inexistente, uma vez que todos os recursos encontrados eram necessários para a alimentação do grupo. Em determinado momento da evolução humana, tornou-se possível observar e compreender a forma pela qual as plantas originavam-se e qual era a relação entre a semente e a existência de novas fontes de alimento. Naquela época, a espécie humana passou a aplicar o conhecimento dessas observações da natureza para plantar o que poderia ser utilizado para sua própria alimentação, surgindo os primórdios da agricultura.

A agricultura permitiu que as populações humanas não mais necessitassem migrar para adquirir novos alimentos e pudessem se estabelecer em regiões mais propícias ao cultivo. Em decorrência desse processo, a espécie humana passou a ter uma relação mais direta com as plantas, num processo de **coevolução**. Na origem da agricultura, a religiosidade provavelmente deve ter tido um papel importante, pois o homem primitivo acreditava que ao "doar" parte dos alimentos para a terra, alguma divindade iria recompensá-lo com mais suprimentos. Em decorrência, também aconteceria um reflexo na qualidade dos alimentos cedidos, uma vez que, quanto melhor fosse a "doação" (sementes plantadas), melhor seria a "recompensa" (colheita). Naquele momento, o homem iniciou um processo de **seleção artificial** para características que acreditava serem melhores, que eram desejáveis, modificando de forma profunda e definitiva o genoma das plantas sob cultivo ao longo das gerações.

DOMESTICAÇÃO: UM PROCESSO DE COEVOLUÇÃO

Ainda que, provavelmente, o processo de escolha das melhores características das plantas, provavelmente, tenha surgido por motivos religiosos, indiretamente, acabou se tornando um fator seletivo determinante. Desde então, a espécie humana vem selecionando plantas que se enquadram melhor em um **ideótipo**. As características selecionadas muitas vezes foram diferentes daquelas mais vantajosas à sobrevivência e propagação das plantas em condições naturais. Nesse sentido, a seleção artificial muitas vezes foi realizada em antagonismo à ação da **seleção natural**, fazendo com que, ao se aproximar do ideótipo estabelecido pela espécie humana, o genoma das plantas cultivadas se distanciasse de seus **parentais selvagens**. A esse processo de modificação genética do genoma de uma espécie, por seleção natural, dá-se o nome de domesticação.

No que se refere às plantas, desde que o homem passou a cultivá-las e, consequentemente, selecioná-las deu-se início ao processo de domesticação. A domesticação é um fenômeno que deve ser interpretado como um processo de coevolução, nesse caso, entre a espécie humana e as plantas cultivadas. A espécie humana selecionou e utilizou as plantas para seu proveito, desse modo elas foram beneficiadas com a disponibilidade regular de água e nutrientes, o controle de pragas e doenças, a eliminação de plantas invasoras e com o auxílio na propagação. A domesticação resulta de um processo cooperativo em que seu ápice se dá na total dependência de uma das espécies envolvidas em relação à outra.

Seleção artificial: processo de seleção conduzido pelo ser humano. Geralmente são realizados cruzamentos controlados que produzem populações segregantes nas quais são selecionados indivíduos ou famílias com características desejáveis.

Ideótipo: forma ideal de planta para determinado ambiente e objetivo de cultivo.

Seleção natural: processo em que pela maior viabilidade e fecundidade de determinados indivíduos, em decorrência da sua maior adaptação a determinada condição ecológica, a frequência de descendentes destes indivíduos é aumentada ao longo das gerações.

Parental selvagem: unidade taxonômica que possui uma relação de parentesco com outra que passou pelo processo de domesticação ou que tem sido utilizada para o cultivo.

Coevolução: evolução simultânea e interdependente, entre duas ou mais espécies, decorrente das interações ecológicas que ocorrem entre elas, fazendo com que a evolução de uma das espécies envolvidas esteja parcialmente dependente da evolução que ocorre na outra espécie.

A DOMESTICAÇÃO REQUER VARIAÇÃO GENÉTICA

O processo de selecionar materiais que melhor se enquadram em um ideótipo somente foi (e é) possível graças à variação genética que existe naturalmente nesses organismos. Essa variação, em conjunto com a variação ambiental (e o efeito conjunto de ambas), é responsável por gerar a diversidade morfológica que pode ser observada em qualquer ser vivo. No que se refere ao processo de domesticação de plantas, a existência de variação permitiu que a espécie humana optasse por utilizar aquelas que mais se aproximassem do ideótipo estabelecido. A espécie humana teve um papel expressivo na evolução de determinadas espécies, por ser naturalmente curioso e ficar atento a características distintas das diferentes espécies de plantas. Modificações morfológicas aberrantes que surgiram ocasionalmente nas plantas, que muitas vezes seriam pouco vantajosas em condições naturais, chamaram a atenção e foram propagadas sob condições de cultivo.

Hábito de crescimento: caráter morfoagronômico, determinado pelo crescimento do ramo principal e pelo florescimento da planta.

A utilização da variação morfológica na busca por um ideótipo a ser cultivado envolveu o acúmulo de características agronomicamente favoráveis denominadas de traços de domesticação que, em conjunto, são denominadas de síndromes de domesticação. Essas síndromes podem variar de espécie para espécie, mas geralmente incluem certos fenótipos como, por exemplo, a perda de dormência, o aumento no tamanho das sementes, o aumento do tamanho dos frutos e o **hábito de crescimento** determinado. Uma espécie interessante para se compreender as síndromes de domesticação é o girassol (*Helianthus annuus*). O parental selvagem do girassol era arbustivo e possuía múltiplas flores pequenas espalhadas por toda a planta. O processo de domesticação fez com que o girassol cultivado viesse a ter um porte mais ereto e uma única flor grande e no ápice do corpo da planta. As mudanças ocorridas na morfologia do girassol apresentam uma forte relação com o ideótipo de uma planta cultivada pelo homem.

COMO SE DÁ A VARIAÇÃO EM NÍVEL DE DNA?

A variação morfológica dos organismos está associada à variação na sequência do DNA. Essa variação pode ocorrer em um ou poucos nucleotídeos ou chegar a alterar a estrutura e o número de cromossomos nas diferentes espécies. Essas variações são chamadas pelos geneticistas de polimorfismos e têm origem em mutações que ocorrem com o processo de divisão celular. Quando estão relacionadas a um evento de alteração morfológica de uma determinada espécie, elas são denominadas de polimorfismos causais.

Em relação ao nucleotídeo, esses polimorfismos podem ser consequência da substituição de um único nucleotídeo por outro – SNP (do inglês: *Single Nucleotide Polymorphism*), ou ainda, pela perda ou ganho de alguns nucleotídeos – InDel (do inglês: *Insertion or Deletion*). Uma representação gráfica da ocorrência de um SNP e de um InDel pode ser observada na Figura 1.

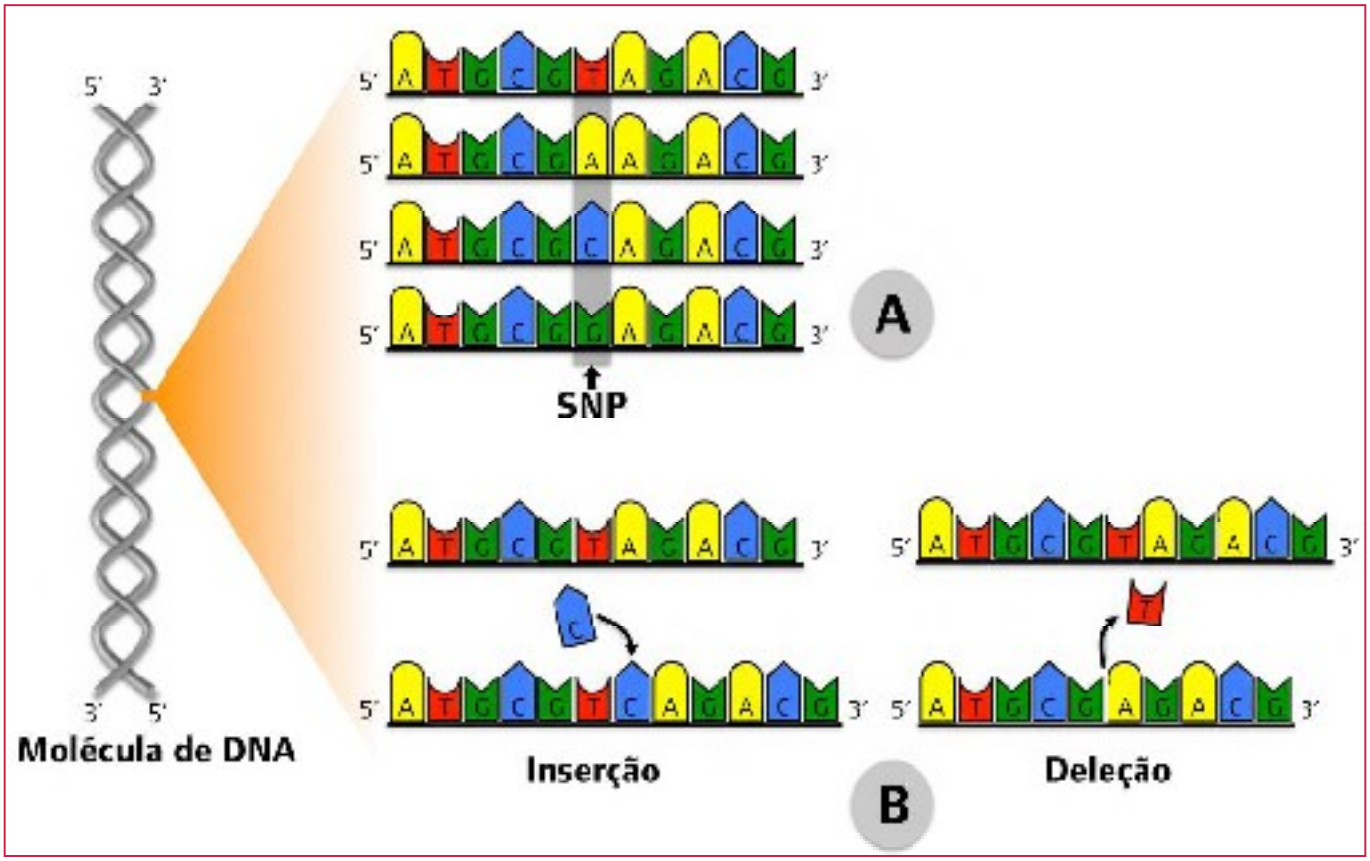


Figura 1.
Representação das mutações que podem ocorrer na molécula de DNA relacionadas a um ou poucos nucleotídeos: A) SNP – substituição de uma base nitrogenada por outra (Timina, Adenina, Citosina ou Guanina, em fundo cinza) e B) InDel – representado pela inserção de uma base C (esquerda) ou pela deleção de uma base T (direita).

O GENE *qSH1* DE ARROZ (*Oryza sativa*)

Dentre as plantas cultivadas, os cereais se destacam quanto ao seu consumo pela espécie humana desde as populações mais primitivas. Dentre as plantas pertencentes a esse grupo (e que são consumidas mundialmente) pode-se citar a aveia, o trigo, o milho, a cevada e o arroz. Estes cereais foram domesticados a partir de espécies de gramíneas silvestres. Provavelmente, um dos passos iniciais na domesticação dessas espécies foi a redução do desprendimento dos grãos quan-

Degrane: queda natural dos grãos de uma planta que ocorre quando se alcança a maturação.

Lócus: local específico no cromossomo onde determinado gene ou alelo é encontrado.

QTLs: Lócus que participam do controle genético de caracteres quantitativos. Identificados pela associação do polimorfismo de marcadores genéticos à variação observada na expressão fenotípica do caractere quantitativo. Corresponde à região do DNA cujo polimorfismo está relacionado à variação de um fenótipo.

Variedades: em agronomia, refere-se às populações, geralmente melhoradas, que diferem entre si em caracteres de importância agronômica.

do maduros (**degrane**). A liberação dos grãos maduros gera uma queda na eficiência do processo de colheita e, consequentemente, é uma característica não desejada para uma planta cultivada.

Muito provavelmente, o início da seleção para a redução do degrane ocorreu de forma inconsciente, pois os grãos que caiam com o amadurecimento não eram colhidos e, consequentemente, não eram utilizados para o plantio. Essa seleção "inconsciente" provocou um aumento da frequência de alelos associados à permanência do grão maduro na planta, ou seja, alelos desfavoráveis ao degrane, em populações de cereais que eram cultivadas pelo homem. A redução do degrane em cereais é um bom exemplo do efeito do processo de domesticação, pois exemplifica as alterações genéticas geradas pelo homem nessas plantas em decorrência da seleção artificial, deixando sinais ou síndromes de domesticação das espécies cultivadas.

Em se tratando do arroz cultivado (*Oryza sativa*), a perda do degrane foi um dos eventos mais importantes que ocorreram durante o processo de domesticação da espécie. Um estudo publicado na revista científica Science em 2006 por Konishi e colaboradores, detectou que cinco regiões no genoma de arroz eram responsáveis por controlar o caráter de degrane da espécie. Essas regiões encontradas no genoma de arroz são chamadas **lócus** de caracteres quantitativos (**QTLs** – do inglês: *Quantitative Trait Loci*). Eles são assim chamados, pois cada uma dessas regiões (lócus) possui uma parcela de contribuição na variação fenotípica do caráter degrane em arroz. Esses QTLs foram identificados em uma população de plantas de arroz derivada do cruzamento entre duas **variedades** de arroz, denominadas *Kasalath* e *Nipponbare*. A variedade *Kasalath* pertence à subespécie de arroz *Oryza sativa* subsp. *indica* e costuma apresentar degrane relativamente forte. Contrariamente, a variedade *Nipponbare*, que pertence à subespécie *Oryza sativa* subsp. *japonica*, normalmente não apresenta essa característica.

Todos os cinco QTLs identificados na população resultante da mistura de *Nipponbare* e *Kasalath* demonstraram contribuir para a redução do degrane em arroz. Eles foram identificados em cinco diferentes cromossomos da população em estudo. Em três QTLs, lo-

calizados nos cromossomos 1, 2 e 5, os alelos de redução de degrane são oriundos da variedade *Nipponbare*. Para os outros dois, localizados nos cromossomos 11 e 12, os alelos de redução de degrane foram identificados na variedade *Kasalath*. Esse resultado, inicialmente, sugeriu que a redução do degrane pode ter ocorrido de forma independente nas subespécies de *O. sativa* subsp. *indica* e *O. sativa* subsp. *japonica*, já que ambas possuem QTLs relacionados com esse caráter.

Um dos QTLs identificados, denominado de gene *qSH1* (*seed shattering*) está localizado no cromossomo 1 e explicou sozinho 68,6% da variação fenotípica total observada para degrane, na população de arroz avaliada. Assim, ele é o principal QTL envolvido no controle do caráter de degrane em arroz. Além disso, foi observado que indivíduos da variedade *Nipponbare* que tinham o *qSH1* proveniente da variedade *Kasalath* apresentavam a formação de uma **zona de abscisão** completa entre o **pedicelo** e a espigueta na base da semente de arroz, ou seja, apresentavam um fenótipo de perda de sementes (degrane).

Zona de abscisão: região especializada na planta, em que mudanças químicas e físicas facilitam a ruptura e liberação de determinadas estruturas (como ocorre, por exemplo, na queda de folhas, i.e. na abscisão foliar).

Pedicelo: estrutura da planta responsável pela sustentação de um determinado órgão e condução de seiva para ele.

Foi realizado um mapeamento genético da região que inclui o gene *qSH1* e nela foi observada um SNP em uma região de 612 pb entre dois dos **marcadores moleculares** (*qSH1-F* e *qSH1-H*), utilizados na identificação de QTLs. Uma predição de genes feita para o local do gene *qSH1* nos genomas das variedades *Nipponbare* e *Kasalath* mostrou não haver distinção entre as **ORFs** na região onde ocorre esse SNP. Entretanto, a uma distância de 12kb desse SNP foi encontrada uma ORF de um gene de arroz **ortólogo** ao gene *REPLUMLESS* (*RPL*) de *Arabidopsis*. O gene *RPL* codifica uma proteína chamada *BELL1-type* e está envolvido na deiscência dos frutos de *Arabidopsis*.

Marcadores moleculares: todo e qualquer fenótipo molecular oriundo de um gene expresso, como no caso de isoenzimas, ou de um segmento específico de DNA (correspondente a regiões expressas ou não do genoma).

ORFs: do inglês *Open Reading Frames*, correspondem às sequências de DNA contidas entre os códons de iniciação e terminação que são lidas pelo maquinário genético para a produção do mRNA.

Regiões ortólogas: Regiões genômicas que compartilham um ancestral comum e são derivadas de um evento de especiação. A função das mesmas é geralmente preservada durante a evolução.

Acredita-se que, assim como o gene *RPL* em *Arabidopsis*, o gene *qSH1* pode ter relação com o desenvolvimento das **espiguetas** e com a formação da camada de abscisão, de modo que a ocorrência de mutações nas ORFs pode ter provocado anomalias durante o desenvolvimento do arroz cultivado. A proteína *BELL1-type* de *Arabidopsis* é um fator de transcrição putativo, ou seja, existem evidências de que essa proteína atue no controle da expressão gênica. As análises realizadas neste estudo mostram que o SNP encontrado pode ter causado a perda da expressão do **mRNA** no gene *qSH1*. Além disso, acredita-se que este SNP poderia ter sobrevivido ao processo de domesticação do arroz.

Espiguetas: unidade básica da inflorescência de uma gramínea (planta que pertence à família Poaceae).

mRNA: molécula de RNA responsável pelo transporte da informação genética do núcleo para os ribossomos, na realização da síntese de proteínas.

O GENE *qSH1* DE ARROZ: CARACTERIZAÇÃO

O gene *qSH1* encontra-se no cromossomo 1 de arroz, apresentando o comprimento de 3893 bases. Este gene possui quatro éxons e três íntrons. Todos os éxons juntos possuem um comprimento total de 1839 bases, sendo o último deles o maior, com 807 bases (Figura 2).

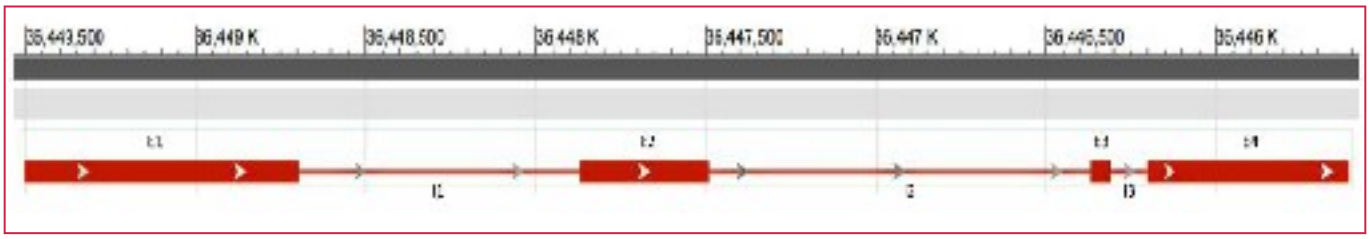


Figura 2.
Representação gráfica do gene *qSH1* em arroz (*Oryza sativa*). A faixa cinza escura representa o cromossomo 1 desta espécie e possibilita a visualização da região em que o gene *qSH1* (em vermelho) se localiza. Os retângulos vermelhos representam os quatro exóns deste gene (E1, E2, E3 e E4) e as linhas entre eles representam os íntrons (I1, I2 e I3) (Fonte: http://www.ncbi.nlm.nih.gov/gene/?term=qSH1).

Um SNP localizado na região 5' do gene, a 12Kb da região de código, está relacionado ao fenótipo degrane. Indivíduos que apresentam a variante T (Timina) exibem um fenótipo de degrane das sementes no arroz enquanto que os que possuem a variante G (Guanina) apresentam um fenótipo de não-degrane (Figura 3).

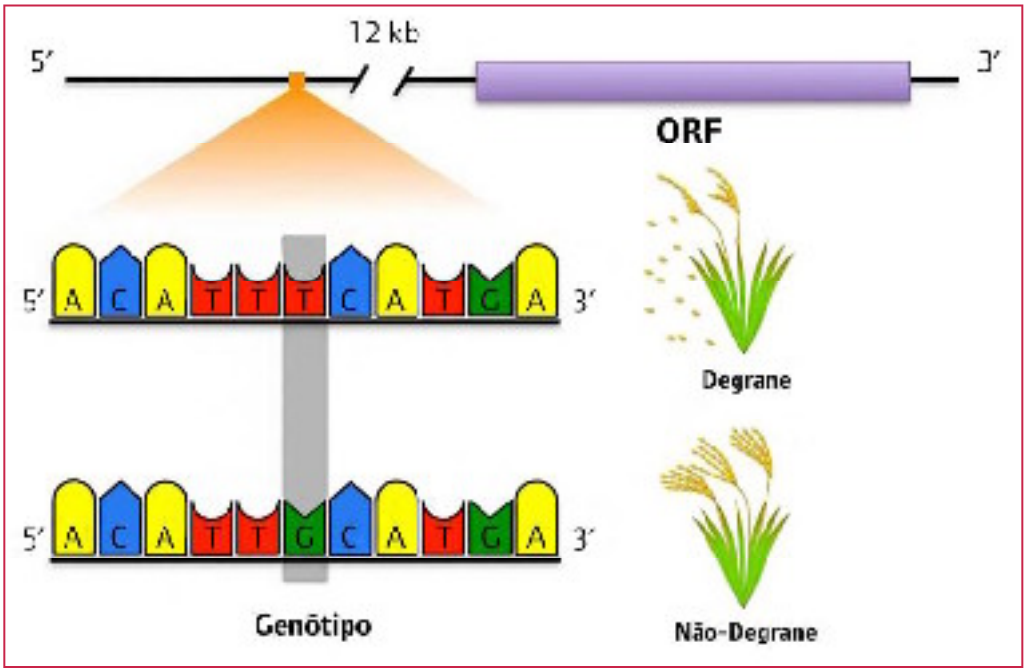


Figura 3.
Representação do SNP no genoma do arroz em que o alelo T está relacionado ao degrane e o alelo G ao não-degrane das sementes. Este SNP ocorre a 12 Kb de distância do ponto de início da transcrição do gene alvo (ORF).

Para verificar como o SNP permaneceu na região do gene *qSH1* durante a domesticação, foram analisadas outras diversas plantas de arroz. Os resultados revelaram que o SNP estava altamente associado ao grau de degrane entre cultivares de *O. sativa* subsp. *japonica*, sugerindo que este grupo havia sido alvo de seleção artificial para o hábito de não-degrane durante a domesticação do arroz. Com relação às cultivares de *O. sativa* subsp. *indica* testadas, estas apresentaram uma alta taxa de degrane, e continham a versão funcional do gene em seus genomas.

Por meio da análise da região do genoma onde se encontra o gene *qSH1*, o SNP identificado pode ser atribuído a uma mutação que ocorreu em populações domesticadas, e não em populações silvestres, da subespécie *O. sativa* subsp. *japonica*. No processo hipotético de evolução do gene *qSH1*, a distribuição do SNP revelou um forte sinal de seleção artificial durante a domesticação do arroz, uma vez que o alelo do SNP está associado à expressão do gene *qSH1*.

PARA SABER MAIS

BARBIERI, R. L. *Origem e Evolução de Plantas Cultivadas*. Embrapa Informação Tecnológica; Pelotas: Embrapa Clima Temperado, 2008.

HARLAN, J. R. *Crops and Man*. Madison, Wisconsin, American Society of Agronomy, 1975.

KONISHI, S.; IZAWA, T.; LIN, S. Y., EBANA, K.; FUKUTA, Y.; SASAKI, T.; YANO, M. An SNP Caused Loss of Seed Shattering During Rice Domestication. *Science*, v. 312, n. 5778, p. 1392-1396, 2006.

Apêndice 2

A dupla hélice: como descobri a estrutura do DNA

Watson, J. A dupla hélice: como descobri a estrutura do DNA. Rio de Janeiro: Ed Zahar, 2013

Rhewter Nunes[1], Mariana Pires de Campos Telles[2]*

[1] Programa de Pós-Graduação em Genética e Melhoramento de Plantas, Escola de Agronomia, Universidade Federal de Goiás (UFG), Goiânia, GO
[2] Escola de Ciências Agrárias e Biológicas, Pontifícia Universidade Católica de Goiás (PUC-GO) e Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás (UFG), Goiânia, GO

* E-mail para correspondência: tellesmpc@gmail.com

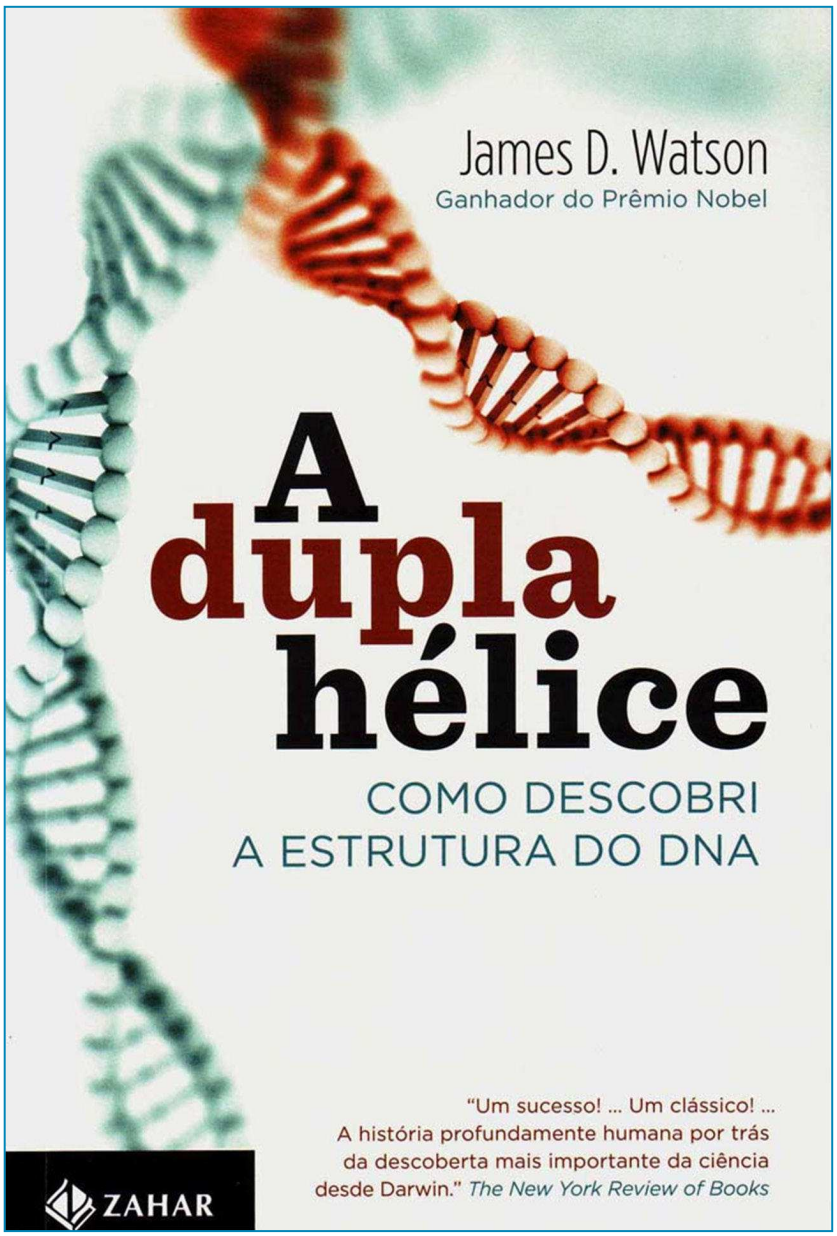
Um dos avanços científicos mais importantes do século XX, a elucidação da estrutura da molécula de DNA, trouxe subsídios importantes para a compreensão de como a informação genética é armazenada e transmitida entre gerações. Em *"A dupla hélice: como descobri a estrutura do DNA"*, James D. Watson faz um relato sobre os acontecimentos que precederam essa descoberta revolucionária. No Brasil, este livro foi publicado pela editora Zahar em 2014, com tradução de Rachel Botelho e revisão técnica de Denise Sasaki.

Neste livro, James Watson apresenta sua versão de como foi a sucessão de acontecimentos que viria a culminar na sua premiação com um Nobel, juntamente com Francis Crick e Maurice Wilkins, em 1962. Durante essa narrativa é praticamente impossível não se contagiar e se inspirar com o jovem Watson, então pesquisador de apenas 24 anos e que tinha uma vontade incessante de deixar sua marca no mundo da ciência. Essa energia e jovialidade pode ser percebida, por exemplo, no trecho *"Era certamente melhor me imaginar ficando famoso do que envelhecendo como um acadêmico reprimido que nunca arriscara uma ideia"* ao relatar sua motivação em estudar o *"segredo da vida"*.

Além de um relato detalhado de como foi, para ele, o processo de estudo e construção do conhecimento necessário para se chegar à descoberta da estrutura do DNA. Durante a leitura é possível ter acesso às angustias profissionais e pessoais de Watson, sua relação, de extrema admiração, com Francis Crick, além de inúmeras observações pessoais e profissionais sobre várias outras pessoas como Maurice Wilkins, Rosalind Franklin, Linus Pauling, seus professores e até mesmo a família de Francis.

A linguagem é bastante acessível, apresenta notas de rodapé com explicações de conceitos científicos importantes e está repleto de

esquemas que ilustram moléculas químicas. Além disso, disponibiliza um encarte com fotos em preto e branco de acontecimentos importantes e da maioria das personalidades descritas por Watson em seus apontamentos. O livro é indicado para qualquer um que deseja conhecer um relato detalhado com o olhar de um dos pesquisadores diretamente envolvidos nessa descoberta tão importante sobre a molécula de DNA. É ainda mais indicado para aqueles que desejam se inspirar em construir uma carreira relacionada à genética.



Por se tratar de um livro relativamente curto (pouco mais de 200 páginas), de fácil compreensão e que apresenta um contexto histórico de conceitos científicos importantes para se entender a estrutura do DNA, ele pode ser utilizado em aulas de biologia do ensino médio, ou ainda, indicado para ingressantes dos cursos de Biologia. Diferentes tópicos do tema apresentados na obra podem permitir a discussão de assuntos importantes sobre ciência em sala de aula. Os relatos históricos do processo de conseguir elucidar a molécula de DNA podem ser utilizados para a discussão de como o processo de produção de conhecimento científico é realizado. O fato desse achado científico ser amparado por resultados de trabalhos anteriores, como os de Chargaff, por exemplo, pode ser utilizado para dialogar com os estudantes quanto ao fato de que as descobertas científicas não são verdades absolutas e estáticas e sim que há um acúmulo de evidências que permitem uma boa resposta para uma determinada questão, num determinado intervalo de tempo.

Ainda nesse sentido, o relato de Watson quanto à idealização inicial de um modelo errado para a molécula de DNA com ligações de magnésio, que foi criticado por Rosalind Franklin posteriormente, pode ser utilizado na discussão de como o "fazer ciência" ocorre de maneira colaborativa e que um problema científico, muitas vezes, precisa ser avaliado por especialistas de diversas áreas e com diferentes habilidades e competências. O próprio Watson menciona em um trecho no capítulo de abertura *"...há uma ignorância generalizada sobre como a ciência é 'feita'. Isso não significa que toda ciência é feita da maneira descrita aqui. Está longe de ser o caso, já que os estilos de pesquisa científica são tão variáveis quanto as personalidades humanas."* As afirmações deste trecho podem levantar uma discussão importante sobre questões, tais como: 1) Como funciona o método científico?; 2) Ainda que diferentes trabalhos utilizem o método científico para gerar evidências, existe uma única forma de se fazer ciência?; 3) Se o produto do conhecimento científico não é necessariamente a "verdade", como esse tipo de conhecimento difere do senso comum?. Todas essas discussões podem ser levantadas em diferentes níveis de profundidade conforme o nível de ensino em que o livro esteja sendo trabalhado.