

---

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS  
(BIOLOGIA VEGETAL)

---

DESENVOLVIMENTO DE RECURSOS GENÔMICOS E TRANSCRICIONAIS EM

*Quillaja brasiliensis*

GIAN CARLOS COSTA BARROS

---

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS  
(BIOLOGIA VEGETAL)

---

DESENVOLVIMENTO DE RECURSOS GENÔMICOS E TRANSCRICIONAIS EM  
*Quillaja brasiliensis*

GIAN CARLOS COSTA BARROS

Dissertação apresentada ao Instituto de Biociências do Câmpus de Rio Claro, Universidade Estadual Paulista, como parte dos requisitos para obtenção do título de Mestre em Ciências Biológicas (Biologia Vegetal).

**Orientador: Douglas Silva Domingues**

Rio Claro – SP  
2022

B277d      Barros, Gian Carlos Costa  
Desenvolvimento de recursos genômicos e transcricionais  
em *Quillaja brasiliensis* / Gian Carlos Costa Barros. -- Rio  
Claro, 2022  
66 p.  
  
Dissertação (mestrado) - Universidade Estadual Paulista  
(Unesp), Instituto de Biociências, Rio Claro  
Orientadora: Douglas Silva Domingues  
  
1. *Quillaja brasiliensis*. 2. saponinas triterpênicas. 3.  
biotecnologia. 4. recursos moleculares. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do  
Instituto de Biociências, Rio Claro. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

**CERTIFICADO DE APROVAÇÃO**

**TÍTULO DA DISSERTAÇÃO:** Desenvolvimento de recursos genômicos e transcricionais em *Quillaja brasiliensis*

**AUTOR:** GIAN CARLOS COSTA BARROS

**ORIENTADOR:** DOUGLAS SILVA DOMINGUES

Aprovado como parte das exigências para obtenção do Título de Mestre em Ciências Biológicas (Biologia Vegetal), área: Biologia Vegetal pela Comissão Examinadora:



Pesquisador Dr. DOUGLAS SILVA DOMINGUES (Participação Virtual)  
Departamento de Botânica / IB UNESP Rio Claro

Prof. Dr. ARTHUR GERMANO FETT-NETO (Participação Virtual)  
Botânica - Instituto de Biociências / Universidade Federal do Rio Grande do Sul

Profa. Dra. MARINA ALVES GAVASSI (Participação Virtual)  
Instituto de Biociências de Rio Claro-IB-UNESP / Rio Claro/SP

Rio Claro, 29 de novembro de 2022

## Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo 2019/15477-3. Para o projeto “Moléculas sinalizadoras na modulação do transcriptoma, fluxo metabólico e produção de saponinas triterpênicas imunoadjuvantes em *Quillaja brasiliensis* Mart. (Quillajaceae)”. Ao Instituto de Biociências de Rio Claro pelo suporte durante o desenvolvimento da minha dissertação; agradecimentos e carinho especial ao grupo de Genômica e Transcricção de plantas do departamento de Botânica, especialmente pela orientação do Prof. Douglas Silva Domingues; aos meus amigos e familiares, por todo o apoio e carinho durante os anos de vivência.

## Resumo

A Família Quillajaceae pertence à ordem Fabales e é composta por duas espécies: *Quillaja saponaria* e *Quillaja brasiliensis*, ambas nativas da América do Sul. Esse gênero é conhecido, do ponto de vista biotecnológico, pela produção de saponinas triterpênicas com uso na indústria farmacêutica como imunoestimulante. Apesar do potencial biotecnológico da espécie, *Q. brasiliensis* não possui dados genômicos ou transcricionais públicos que permitiriam maiores avanços no desenvolvimento de ferramentas biotecnológicas. Nesse contexto, o presente estudo teve por objetivo o desenvolvimento de recursos genômicos e transcricionais para *Quillaja brasiliensis*. Foram desenvolvidos recursos genômicos por meio da plataforma Illumina de sequenciamento, e recursos transcricionais por meio das plataformas PacBio e Illumina, com avaliação de folhas e culturas celulares. A partir dos dados genômicos, foi montado o genoma cloroplastidial de *Q. brasiliensis*, que apresenta 160528 pares de base. Com ele, foram realizados estudos filogenéticos comparativos, incluindo-se dados plastidiais de espécies das famílias Quillajaceae, Surianaceae e Polygalaceae. Com base nos recursos transcricionais PacBio, foi montado um transcriptoma de referência para *Q. brasiliensis*. Com base nessa referência, identificamos por sequenciamento Illumina 588 genes com estabilidade potencial para uso como normalizadores, dos quais 13 foram selecionados para testes iniciais por PCR. Três deles foram selecionados para teste em RT-qPCR: *Heat shock protein Hsp90* (HSP), *Cyclophilin-like domain* (CYP) e *Ubiquitin* (UBQ). Estes genes tiveram sua estabilidade em folhas, caule e raízes. HSP foi o gene que apresentou os resultados mais consistentes de estabilidade considerando-se os três tecidos avaliados. Nas análises individuais em cada tecido, CYP foi o gene normalizador mais bem ranqueado para folha e raiz e HSP o mais bem ranqueado em caule. Com isso, o presente estudo estabelece recursos moleculares iniciais em *Q. brasiliensis*, que podem ser utilizados tanto para o estudo do genoma e análises filogenéticas bem como para a descoberta de genes candidatos à produção de saponinas triterpênicas.

Palavras-chave: *Quillaja brasiliensis*, saponinas triterpênicas, biotecnologia, recursos moleculares.

## Abstract

The Quillajaceae family belongs to the order Fabales and it is composed by two species: *Quillaja saponaria* and *Quillaja brasiliensis*, both native to South America. This genus is known, from a biotechnological point of view, by the production of triterpene saponins used in the pharmaceutical industry as an immunostimulant. Despite the species biotechnological potential, *Q. brasiliensis* does not have public genomic or transcriptional data that would allow further advances in the development of biotechnological tools. In this context, the present study aimed to develop genomic and transcriptional resources for *Quillaja brasiliensis*. Genomic resources were developed through the Illumina sequencing platform, and transcriptional resources through the PacBio and Illumina platforms, with evaluation of leaves and cell cultures. From the genomic data, the chloroplastidial genome of *Q. brasiliensis* was assembled, which has 160,528 base pairs. We then carried out comparative phylogenetic studies, including plastid data from species of the Quillajaceae, Surianaceae, and Polygalaceae families. Based on PacBio cDNA sequencing, a reference transcriptome for *Q. brasiliensis* was assembled. Based on this reference, we identified by Illumina sequencing 588 genes with potential stability for use as normalizers, of which 13 were selected for PCR testing. Three of them were selected for tests in RT-qPCR: Heat shock protein Hsp90 (*HSP*), Cyclophilin-like domain (*CYP*) and Ubiquitin (*UBQ*). These genes had their stability evaluated in leaves, stems and roots. *HSP* was the gene that presented the most consistent stability results considering the three evaluated tissues. In the individual analyzes in each tissue, *CYP* was the highest ranked normalizing gene for leaf and root and *HSP* was the highest ranked in stem. Thus, the present study establishes initial molecular resources in *Q. brasiliensis*, which can be used both for the study of the genome and phylogenetic analyzes as well as for the discovery of candidate genes for the production of triterpene saponins.

**Key words:** *Quillaja brasiliensis*, triterpenic saponins, biotechnology, molecular resources.

## Sumário

|                                                                                                                              |    |
|------------------------------------------------------------------------------------------------------------------------------|----|
| Resumo .....                                                                                                                 | 4  |
| Abstract .....                                                                                                               | 5  |
| 1. Introdução .....                                                                                                          | 8  |
| 1.1. Genoma dos cloroplastos.....                                                                                            | 12 |
| 1.2. Sequenciamento Illumina .....                                                                                           | 13 |
| 1.3. Sequenciamento Iso-seq.....                                                                                             | 14 |
| 1.4. Genes Normalizadores.....                                                                                               | 15 |
| 2. Objetivo .....                                                                                                            | 16 |
| The complete chloroplast genome of <i>Quillaja brasiliensis</i> and comparative analysis<br>in Quillajaceae and Fabales..... | 19 |
| Introduction.....                                                                                                            | 19 |
| Material and methods .....                                                                                                   | 19 |
| Plant material, DNA extraction and sequencing.....                                                                           | 19 |
| Genome assembling, annotation and SSR analysis .....                                                                         | 20 |
| Sequence divergence analysis .....                                                                                           | 20 |
| Phylogenetic analysis .....                                                                                                  | 21 |
| Results.....                                                                                                                 | 21 |
| <i>Q. brasiliensis</i> and <i>Q. saponaria</i> assembly .....                                                                | 21 |
| Overall organization and general features of the plastomes .....                                                             | 21 |
| Discussion .....                                                                                                             | 31 |
| Changes in the content of plastomes: gene gain/loss .....                                                                    | 31 |
| Variable regions were identified based on plastome sequences.....                                                            | 32 |
| Phylogenomic analysis and relationship among species.....                                                                    | 32 |
| Conclusion .....                                                                                                             | 33 |
| Reference .....                                                                                                              | 34 |
| Análises transcricionais em <i>Quillaja brasiliensis</i> : identificação de genes<br>normalizadores.....                     | 38 |

|                                                                     |    |
|---------------------------------------------------------------------|----|
| Introdução .....                                                    | 38 |
| Material e metodos.....                                             | 40 |
| Material vegetal e extração de RNA .....                            | 39 |
| RNA-seq: sequenciamento Iso-seq e Illumina .....                    | 40 |
| Montagem dos transcritos e quantificação de expressão .....         | 41 |
| Seleção de genes normalizadores e desenho de <i>primers</i> .....   | 42 |
| Síntese de cDNA e PCR convencional .....                            | 42 |
| RT-qPCR .....                                                       | 43 |
| Estabilidade de genes de referência.....                            | 43 |
| Resultados e discussão.....                                         | 44 |
| Sequenciamento Iso-seq: estatísticas gerais.....                    | 44 |
| Normalizadores e análises transcricionais .....                     | 47 |
| Identificação de candidatos a genes normalizadores para RT-qPCR.... | 47 |
| Considerações finais.....                                           | 56 |
| Referências .....                                                   | 56 |

## 1. Introdução

A Família Quillajaceae pertence à ordem Fabales e é composta apenas por duas espécies: *Quillaja brasiliensis* (St. Hil.et Tul.) (Atualmente, *Q. lancifolia*) e *Quillaja saponaria*. A espécie *Quillaja brasiliensis* (Figura 1) é nativa do nordeste da Argentina, sul do Brasil e Uruguai. *Q. saponaria* tem sua ocorrência descrita no Chile



Figura 1: Espécie *Quillaja brasiliensis* (Elaboração Própria).

central (MARTICORENA & MARCHESI 2008, LUEBERT 2013).

Economicamente, esse gênero está relacionado à produção de saponinas triterpênicas. Em espécies angiospermas dicotiledôneas, estas moléculas são usualmente compostas de um triterpeno - terpenos de 30 carbonos, derivados de seis moléculas de isopreno, provenientes do metabolismo citoplasmático de terpenoides (LIAO et al., 2016) - hidrofóbico, muitas vezes chamado de sapogenina.

Além disso, as saponinas contêm de quatro a cinco anéis de agliconas hidrofóbicas. A partir de reações enzimáticas mediadas por enzimas do tipo glicosiltransferases, grupos de açúcar são adicionados em posições específicas dos anéis, bem como podem também ocorrer oxidações mediadas por enzimas P450 monooxigenases (SEKI et al., 2015). Essas agliconas apresentam aspecto apolar e açúcares, que em contato com solução aquosa apresentam um comportamento próximo ao sabão (BISWAS & DWIVEDI 2019; MAGEDANS et al., 2019). Dessa forma, as sapogeninas são agliconas usualmente “decoradas” com sacarídeos que

vão compor a parte hidrofílica da molécula (RIBEIRO et al., 2020; MAGEDANS et al., 2021).

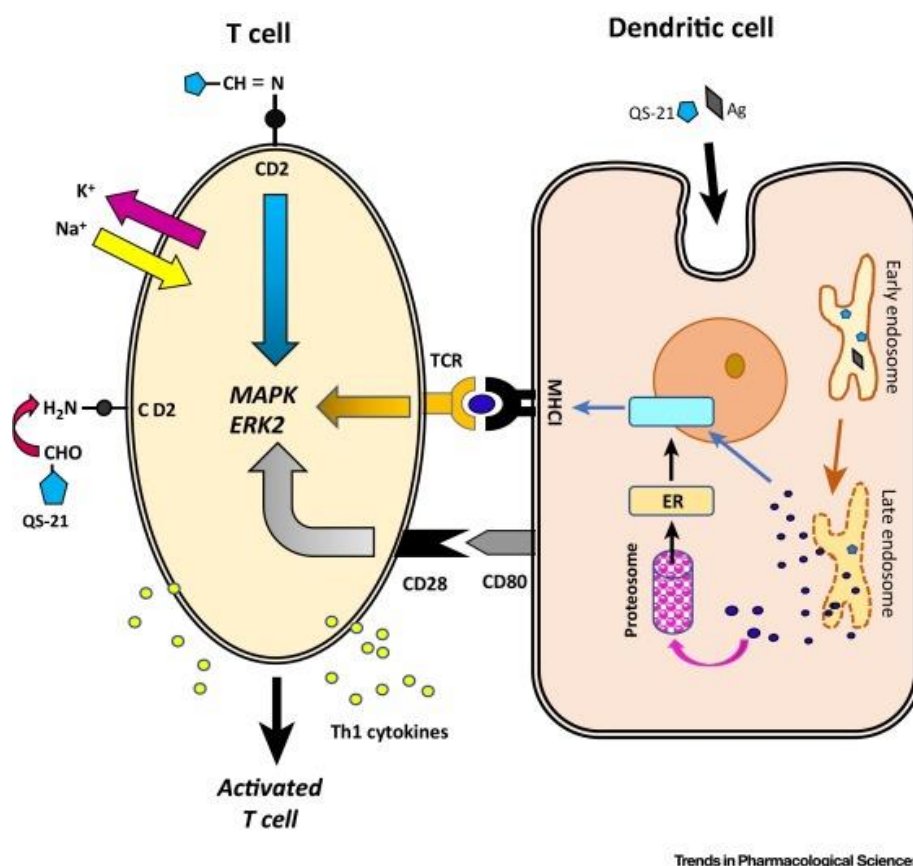
Do ponto de vista ecológico, estes compostos fornecem proteção devido ao seu sabor amargo e adstringente, que impede animais herbívoros de consumirem a planta. As saponinas também servem como substâncias de proteção devido à sua toxicidade para animais monogástricos e insetos herbívoros (RIBEIRO et al., 2020).

As saponinas são moléculas altamente versáteis e apresentam aplicações em diversas atividades industriais. No setor farmacêutico, são relevantes por suas atividades como imunoadjuvantes (promovendo respostas imunes em formulações de vacinas veterinárias e humanas), entre outras propriedades farmacêuticas. Elas também são produtos relevantes na indústria alimentícia. Alguns de seus principais usos são como agentes espumantes em bebidas e emulsificantes em alimentos processados, incluindo encapsulamento e estabilização de compostos bioativos (MAGEDANS et al., 2021).

As saponinas são base de medicamentos expectorante, anti-inflamatório, hipocolesterolêmico e hemolítico. O extrato patenteado Q-Naturale® foi o composto, oriundo de *Q. saponária*, mais estudado no que diz respeito a sua ação imunoadjuvante e emulsificante (DE FARIA et al., 2017; ALIMENTARIUS 2018; FLECK et al., 2019.). Outros compostos de saponinas, como a frações de extratos da casca de *Quillaja saponaria*, exemplo da QS-21, foram testados como imunoestimulante (MARCIANI, 2018).

As saponinas de *Q. brasiliensis*, assim como saponinas de *Q. saponaria* formaram complexos imunoestimulantes (ISCOMs) (CIBULSKI et al., 2016), garantindo para as saponinas de *Q. brasiliensis* uma capacidade de estímulo do sistema imunológico

A ação imunoadjuvante de saponinas de *Quillaja* ocorre de duas maneiras. Em células T, o grupo aldeído se une a QS-21 e desencadeia uma cascata molecular auxílio de quinase ativada por mitógeno (MAP) (ERK2) e em conjunto com transportadores de K<sup>+</sup> e Na<sup>+</sup> resultado na secreção de citocinas Th1. Já em células dendríticas as saponinas agem como facilitadoras da liberação da vesícula endossomal, garantindo uma resposta rápida dessas células e a comunicação imediata com células T medidas por apresentação cruzada do antígeno (MARCIANI 2018). Estes eventos estão descritos na figura 3.



Trends in Pharmacological Sciences

Figura 3: Ação das saponinas em células T e células dendríticas para resposta imunológica. Fonte: MARCIANI 2018.

Nos últimos anos, diversos estudos caracterizaram bioquimicamente as saponinas oriundas da representante brasileira *Quillaja brasiliensis* com potencial imunológico e identificaram características bioquímicas semelhantes as saponinas presentes em *Q. saponaria* (FLECK et al., 2019).

Em estudos de avaliação da capacidade imunoadjuvante de *Q. brasiliensis*, a fração QB-80 foi utilizada para síntese do neoadjuvante IMXQB-80 usado em vacinas experimentais inativadas contra o vírus Zika. Estudos com ratos submetidos a esse neoadjuvante indicaram um aumento significativo de imunoglobulinas IgG e outros subtipos relacionados ao Zika vírus (IgG1, IgG2b, IgG2c), assim como anticorpos na comparação com testes de vacinas sem adjuvante (CIBULSKI et al., 2021).

A espécie *Q. saponaria* é a mais explorada comercialmente para aplicabilidade na indústria alimentícia, de cosméticos e farmacêutica (GONZALEZ &

SÖRENSEN 2020; MA et al., 2020; DAI et al., 2020). As saponinas são extraídas da casca dessa árvore, demandando uma grande quantidade de material vegetal. Assim, as pesquisas recentes investigam formas alternativas de obtenção deste composto (REICHERT et al., 2019).

Dessa forma, a obtenção de saponinas de *Q. brasiliensis* tem maior potencial de sustentabilidade que as de *Q. saponaria*, uma vez que as saponinas estão em maior concentração na folha. Em função desse maior potencial de uso sustentável, o processo de indução da produção de saponinas em *Quillaja brasiliensis* tem sido estudado. Por exemplo, folhas sob efeito de luz UV-C produzem mais saponinas bioativas (YENDO et al., 2015). Outro caminho para o aumento da produção de saponinas bioativas seria por meio de extração destes compostos em culturas celulares (BISWAS E DWIVEDI, 2019).

Como o conhecimento de agentes e condições que estimulam o acúmulo de saponinas é de grande interesse e vários estudos têm sido realizados com esse objetivo. A elucidação das etapas de sua biossíntese em nível genômico e transcricional é um passo crucial nesse sentido.

Uma drástica diminuição dos custos de sequenciamento nos últimos anos tem possibilitado um maior número de projetos de sequenciamento de larga escala (KUMAR et al., 2019). Novos projetos surgiram para a geração de dados genômicos - como o The Earth Biogenome Project (EBP) (<https://www.earthbiogenome.org/>) que pretende contribuir para as diversas áreas do conhecimento como a conservação de espécies ameaçadas.

Em função da proliferação de dados de sequenciamento, quase 4000 genomas cloroplastidiais estão depositados no NCBI (DOBROGOJSKI et al., 2020). Análises do genoma dos cloroplastos obtidos a partir do sequenciamento de nova geração trouxeram maior entendimento sobre relações evolutivas dentro de grupos vegetais além de possibilitar aplicações biotecnológicas como desenvolvimento de vacinas e biofármacos (DANIELL et al., 2016).

Os avanços nas técnicas de sequenciamento também possibilitaram um desenvolvimento acelerado nas análises de RNA. Hoje o RNAseq se transformou em rotina nos trabalhos de biologia molecular. Com ele conseguimos identificar aspectos biológicos do RNA, compreender funcionalmente genes e genomas além de investigar a regulação de genes em diversos organismos (STARK et al., 2019).

No momento, a despeito de seu grande potencial biotecnológico, o desenvolvimento de recursos moleculares para a descoberta de vias metabólicas e compreensão de relações filogenéticas em Quillajaceae é relativamente incipiente. O único estudo público de transcriptoma em Quillajaceae foi o realizado na iniciativa “1kp transcriptomes” (One Thousand Plant Transcriptomes Initiative, 2019), no qual foi sequenciado o transcriptoma foliar de *Q. saponaria* pela técnica de sequenciamento Illumina. Com esta mesma técnica, foi obtido o sequenciamento do genoma de cloroplasto de *Q. saponaria* (Vizoso et al., 2019). Já para *Q. brasiliensis*, não há no momento dados públicos de sequenciamento.

### **1.1. Genoma dos cloroplastos**

No aspecto metabólico, o cloroplasto pode ser considerado a organela mais importante presente nas plantas superiores. O mecanismo bioquímico e biofísico realizado por essa organela está diretamente relacionada a sua função no uso da radiação solar para gerar precursores na geração de energia (Helmut Kirchhoff 2019).

A função dos cloroplastos não se resume ao processo de aproveitamento da radiação solar para a síntese de açúcares para o metabolismo vegetal, mas também estão relacionados a sintetizam fitohormônios, ácidos graxos, aminoácidos além de outros metabólitos secundários, assim atuando diretamente nos aspectos do desenvolvimento vegetal, respostas ao ambiente e a imunidade vegetal (LITTLEJOHN et al., 2021).

As características funcionais e dos cloroplastos estão relacionadas ao seu genoma. Resultado de um processo ancestral de endossimbiose entre organismos eucariotos e uma cianobactéria. Os cloroplastos possuem genoma próprio e altamente diversificado.

A organização genômica dos cloroplastos é bastante variada, podendo apresentar conformações circulares ou linear de DNA, variando de 15.553 pares de bases (pb) em *Asarum minus* a 521.168 pb em *Floydiella terrestris* (BROUARD et al., 2010). Proteínas presentes nos cloroplastos nem sempre são sintetizadas pelo núcleo das células, podendo variar de 0 a 315 genes codificados pelo DNA cloroplastidial. No geral, o genoma cloroplastidial (cpDNA) está organizado em dois fragmentos idênticos chamados repetições invertidas (IR) separadas por uma seção

longa de cópia única (LSC) e uma seção curta de cópia única (SSC) (DOBROGOJSKI et al., 2020).

O sequenciamento de cloroplastos trouxe uma maior compreensão de diversos aspectos biológicos, como a o maquinário transcricional celular e ainda aspectos evolutivos (SANITÁ LIMA et al., 2016). Mesmo com os avanços nas técnicas de sequenciamento de nova geração os cloroplastos têm ficado em segundo plano na comparação com o sequenciamento de mitocôndrias por exemplo (SMITH 2018).

Devido às suas características únicas como a presença de genes altamente conservados e a herança uniparental, o genoma cloroplastidial é fundamental para a reconstrução filogenética de plantas (FANG et al 2017). As filogenias que utilizam os cloroplastos sustentam a monofilia de grandes grupos, como por exemplo a de briófitas (ZHONG et al., 2014) Ou ainda de grupos de importância econômica como o grupo das fabales (AYGOREN ULUER et al., 2020). Mas para que as análises filogenéticas que utilizam dados genômicos de cloroplasto ocorra, é preciso conhecer estes genomas por completo nas diferentes linhagens. Um exemplo é a família foco deste estudo, Quillajaceae, que possui apenas um genoma cloroplastidial reportado de *Q. saponaria* (VIZOSO et al., 2019), mas não existem até o momento trabalhos de sequenciamento com a espécie *Quillaja brasiliensis*.

## **1.2. Sequenciamento Illumina**

A plataforma Illumina foi de extrema importância para o desenvolvimento de estudos relacionados ao sequenciamento de genomas e transcriptomas por trazer rapidez e diminuir custos da técnica de sequenciamento. Este tipo de sequenciamento é baseado na hibridização do DNA a oligonucleotídeos presentes em uma plataforma sólida (lâmina de vidro). diferente da técnica de sanger baseada na amplificação por PCR(VAN DIJK et al., 2018).

O DNA é fragmentado e adaptadores são adicionados em suas extremidades para que elas sejam hibridizadas na lâmina de vidro, em seguida o fragmento se dobra formando pontes. Rodadas sucessivas de síntese de fitas complementares são realizadas nas “pontes” formadas. A adição de nucleotídeos marcados faz com que a fluorescência seja detectada e assim as sequências são elucidadas (KUMAR et al., 2019). Os dados de saída do sequenciamento illumina variam de 1,2GB a 360GB dependendo do número de leituras realizadas para uma dada amostra. Na

plataforma Illumina, as leituras apresentam tamanho entre 150 e 300 bp (SLATKO et al 2018).

O método de sequenciamento utilizado pela plataforma Illumina tem como maior limitação o tamanho de sequência gerado, que vem sendo superado por novas técnicas como a plataforma PacBio e Oxford Nanopore. No sequenciamento illumina, assim como sequenciamento de DNA é possível realizar o sequenciamento do RNA a partir da síntese de um DNA complementar cDNA. O conjunto total de RNA expresso pelas células um organismo é chamado de transcriptoma. Em resumo, a técnica é baseada na fragmentação do RNA gerando leituras curtas assim essas leituras são montadas a partir de ferramentas de Bioinformática (STARK et al., 2019).

O desenvolvimento dessa tecnologia trouxe uma expansão aos estudos funcionais de aspectos biológicos como a regulação e descoberta de genes (STARK et al., 2019). O desenvolvimento dos estudos relacionados ao sequenciamento de RNA trouxe, especificamente para as plantas, possibilidades aos estudos evolutivos e de melhoramento (LE NGUYEN et al., 2019).

Com estes dados, podem ser descobertos genes relacionados à síntese de compostos de interesse em plantas, como genes relacionados à produção de terpenoides em cúrcuma (ANNANDURAI et al., 2013) e pitanga (GUZMAN et al., 2014), bem como a coloração de frutos em aração (VETÖ et al., 2020).

### **1.3. Sequenciamento Iso-seq**

Como uma forma de melhorar o rendimento dos métodos de sequenciamento, um novo método surgiu com a possibilidade de leitura de uma molécula inteira de RNA sem a obrigatoriedade da fragmentação das sequências polipeptídicas para a leitura. Isso possibilitou a identificação de isoformas resultantes dos *splicing* alternativo além de reduzir os erros de montagem (STARK et al., 2019). Este método foi lançado em 2011 pela empresa Pacific Biosciences (PacBio). Nele, a polimerização é realizada em moléculas circulares, diferente dos *reads* fragmentados e fixados em plataforma sólida no Illumina. As moléculas circulares ficam dispostas em poços com uma única polimerase que realiza a amplificação e consecutiva leitura por fluoróforos (KUMAR et al., 2019).

Para os trabalhos de RNA-seq, a tecnologia *long-read* possibilita a identificação clara de RNAs e suas isoformas, possibilitando um maior entendimento

do perfil transcricional da planta uma vez que os erros relacionados a montagem das sequências são menos frequentes (DIJK et al., 2018). Além disso, a tecnologia *long-read* pode contribuir em estudos voltados à descoberta de genes relacionados à síntese de metabólitos especializados, como em espécies utilizadas na medicina chinesa (CUI et al., 2021).

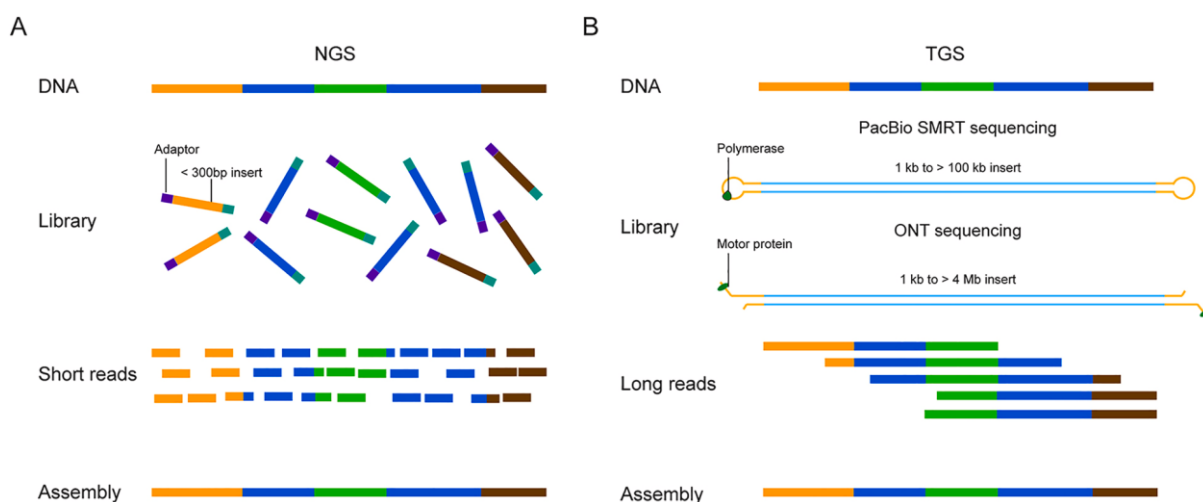


Figura 2: Métodos de leitura curta, equenciamento de alto rendimento(A) e leitura longa TGS (*Long reads*)(B). Destaque para menos sobreposições de reads durante a montagem utilizando *long reads* em comparação com *short reads*. Fonte: CHEN et al., (2021)

Assim, trabalhos que utilizam de métodos de RNA-seq combinados, Illumina e long-read, podem contribuir para a identificação de aspectos metabólicos e funcionais de compostos como os terpenos. Ou ainda possibilitar a construção de transcriptomas completos de forma rápida e econômica de espécies, sem um genoma de referência. Isso foi realizado na espécie de algodão selvagem *Gossypium australe* (FENG et al., 2019) e na espécie da medicina tradicional chinesa *Zanthoxylum armatum* (LIU et al., 2021).

#### 1.4. Genes Normalizadores

Para a validação da expressão diferencial ou identificação de genes chaves para vias metabólicas de interesse, é preciso a identificação de genes considerados normalizadores, ou seja, genes que independente da condição experimental apresentam o mesmo padrão transcricional. Os experimentos de RT-qPCR dependem primariamente de uma extração de RNA, no entanto existe variação no conteúdo total de RNA presentes em uma amostra após a extração, assim o gene de

referência serviria como um calibrador e sua expressão será constante durante os ciclos de amplificação (ZHOU et al., 2017; KOZERA et al., 2013).

Para que o gene normalizador apresente essa característica durante as reações de RT-qPCR é preciso que ele seja previamente selecionado em estudos que caracterizam seu perfil transcricional e os dados obtidos sejam executados para identificar a estabilidade desse gene.

A estabilidade de um normalizador depende de algumas características, são elas; o gene deve ser expresso em diferentes tecidos; o gene precisa ser identificado em diferentes fases de crescimento do organismo de estudo e ser expresso em diferentes condições de tratamento (HELLEMANS et al., 2014). Assim para estudos comparativos, no qual mais de um tratamento é avaliado por RT-PCR quantitativo em tempo real (RT-qPCR), bem como em experimentos que avaliam a expressão de genes em condições diversas, o nível de mRNA do gene de referência é utilizado como calibrador para as condições da análise transcricional (JOSEPH et al., 2018).

Para tanto, é preciso identificar qual gene apresenta características específicas de um normalizador. Para espécies vegetais que não apresentam genes normalizadores conhecidos, é necessária uma avaliação da estabilidade de genes que possam ser utilizados como normalizadores, idealmente comparando tratamentos (JOSEPH et al., 2018). A identificação dos genes mais estáveis, no geral, ocorre por meio da comparação dos valores médios de CT exibidos ou variações desses dados. A plataforma RefFinder é um exemplo amplamente empregado em diversos trabalhos com normalizadores (AIMAS et al., 2018; LI et al., 2021; SARWAR et al., 2020; SHUKLA et al., 2019). Esta plataforma utiliza quatro algoritmos que individualmente indicam a partir dos valores de referência qual o gene mais estável para normalizadores, são elas BestKeeper, NormFinder, Genorm e o método comparativo delta-CT (VANDESOMPELE et al., 2002; ANDERSON et al. 2004; PFAFFL et al., 2004; SILVER et al., 2006).

## **2. Objetivo**

O presente estudo tem como objetivo geral obter recursos genômicos e de transcriptoma da espécie *Quillaja brasiliensis*. Especificamente foram realizados:

- A montagem *de novo* e a anotação do genoma de cloroplasto da espécie;
- Uma análise comparativa do genoma de cloroplastos de espécies da família Quillajaceae;

- Sequenciamento do transcriptoma de folhas e culturas celulares de *Q. brasiliensis*;
- Identificação de genes normalizadores para folhas, caules e raízes de *Q. brasiliensis*.

Para apresentar o alcance desses objetivos, os conteúdos desta dissertação foram divididos em duas partes. No primeiro é apresentado o sequenciamento e análise do genoma plastidial de *Q. brasiliensis*. Na segunda parte apresentamos o desenvolvimento de um transcriptoma de referência para a espécie.

## Capítulo 1

### The complete chloroplast genome of *Quillaja brasiliensis* and comparative analysis in Quillajaceae and Fabales

Gian C. C. Barros<sup>1,2</sup>, Alan P. R. Lorenzetti<sup>1</sup>, Bruno C. Rossini<sup>4</sup>, Yve V. S. Magedans<sup>5</sup>, Celso Luis Marino<sup>4</sup>, Arthur Fett-Neto<sup>5</sup>, Douglas Silva Domingues<sup>1,2</sup>

<sup>1</sup>Group of Genomics and Transcriptomes in Plants, Instituto de Biociências, São Paulo State University, UNESP, Rio Claro, SP, Brazil

<sup>2</sup>Department of Genetics, Escola Superior de Agricultura “Luiz de Queiroz”, University of São Paulo, USP, Piracicaba, SP, Brazil:

<sup>4</sup>Institute of Biotechnology, São Paulo State University, UNESP, Botucatu, SP, Brazil

<sup>5</sup>Center for Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

#### Introduction

Quillajaceae is a family of the order Fabales composed of only two species: *Quillaja brasiliensis* and *Quillaja saponaria* (CHIESA et al 2021). *Quillaja brasiliensis* is native to northeastern Argentina, southern Brazil and Uruguay and *Q. saponaria* has its occurrence described in central Chile (MARTICORENA & MARCHESI 2008, LUEBERT 2013).

Quillajaceae is known to produce saponins, compounds from the specialized metabolism of these plants with bioactive properties (DA SILVA MAGEDANS et al., 2021). Barks and wood of *Q. saponaria* and leaves of *Q. brasiliensis* are the main sources of *Quillaja* saponins. Saponins present in *Quillaja* are mainly used in vaccine adjuvants, due to their strong immunoadjuvant potential, and they have been also used in vaccines against SARS-CoV-2 (WANG, 2021; KUMAR et al., 2022).

Chloroplasts are important metabolic centers in cells, producing carbohydrates and releasing oxygen, through the process of photosynthesis (DANIELL et al., 2016). sequencing has brought a greater understanding of several biological aspects, such as cellular transcriptional machinery and evolutionary aspects (SANITÁ LIMA et al., 2016). With advances in high-throughput sequencing (HTS) technologies, chloroplasts genomes and genomic data have become a good source of information for phylogenetic analyses (DANIELL et al., 2016). The genomic general organization of the chloroplast is based on two single copy regions, small single-copy (SSC) and

large single-copy (LSC), and two inverted repeat regions (IRs; XU et al., 2015). Genome size of chloroplasts is between 100-220 kb, ranging the number of protein-coding genes between 100 to 130 (DANIELL et al., 2016).

Within the Fabales order (four families - Leguminosae, Polygalaceae, Surianaceae, and Quillajaceae), evolutionary studies are concentrated on the legume family Fabaceae due to the large number of species and their economic importance. Recent analyses using chloroplast sequences have indicated a split of the group into 6 subfamilies (ZHAO et al., 2021).

However, other families of Fabales order are particularly neglected in terms of comparative chloroplast genomes. The recent publications of chloroplast genome sequencing for *Quillaja saponaria* (Quillajaceae, VIZOSO et al., 2019) *Suriana maritima* (Surianaceae, LAI et al., 2019) and *Polygala tenuifolia* (Polygalaceae, LEE et al., 2021) now allows comparative studies of chloroplast genomes within Fabales families besides Fabaceae.

In this study, we completely sequenced and compared the chloroplast genome of *Q. brasiliensis* using high-coverage Illumina sequencing and reassembled the chloroplast of *Q. saponaria* using public data. We also analyzed structural genomic characteristics of *Quillaja* chloroplasts, identifying SSR loci to provide resources for later studies in population genetic structure and phylogeography of Quillajaceae. Moreover, we evaluated presence/absence among coding genes as well as analyzed the phylogenetic relationship of Surianaceae, Polygalaceae and Quillajaceae in Fabales using complete chloroplast genome sequences. This is the first comparative analysis of chloroplast genomes in Fabales outside the Fabaceae family.

## **Material and methods**

### **Plant material, DNA extraction and sequencing**

Leaf samples of *Q. brasiliensis* were obtained from germinated seeds that were collected from mature trees in Canguçu, RS, Brazil (31° 23'42"S, 52°40'32"W). Seedlings were kept under white fluorescent lamps at 25 ± 2 °C on a 16/8 h day/night cycle and leaves were collected 11 months after planting. The voucher specimen was deposited at the Herbarium of Federal University of Rio Grande do Sul (ICN 142953). DNA extraction followed the protocol of Healey et al (2014) adapted for recalcitrant

plants. Genomic DNA was sequenced using the Illumina NovaSeq platform at Novogene (<https://en.novogene.com/>) resulting in 165,188,641 paired reads of 150pb.

### **Genome assembling, annotation and SSR analysis**

For the *de novo* assembly of *Q. brasiliensis* and *Q. saponaria* chloroplasts, we used getOrganelle v1.7.5 (JIN et al., 2020). The assembly was performed using the kmer set at 15. Gene annotation was performed using the Geseq-clorobox tool (<https://chlorobox.mpimp-golm.mpg.de/Alternative-Tools.html>) (TILLICH et al., 2017) and CPGAVAS (Shi et al., 2019). PERF was used for SSR analysis (Avvaru et al. 2018). We analyzed mono to hexanucleotides with a minimum of ten repeats for mononucleotides, four for dinucleotides, and three for other motifs.

### **Sequence divergence analysis**

To analyze selection within coding genes, we calculated  $\omega = dN/dS$ , that is, the ratio of non-synonymous to synonymous substitutions. Selection analysis was performed on the MEGAX platform (Kumar et al., 2018) using coding gene sequences of *Prunus persica* (Rosaceae - outgroup), *Cadellia pentastylis* (Surianaceae), *Suriana maritima* (Surianaceae), *Quillaja saponaria* (Quillajaceae), *Quillaja brasiliensis* (Quillajaceae), *Epirixanthes elongata* (Polygalaceae), *Polygala tenuifolia* (Polygalaceae), *Polygala sibirica* (Polygalaceae), *Polygala japonica* (Polygalaceae), *Polygala kerensium* (Polygalaceae), *Salomonina cantoniensis* (Polygalaceae), *Polygala fallax* (Polygalaceae) e *Polygala arillata* (Polygalaceae) chloroplasts.

The complete sequences of *Q. brasiliensis* (Quillajaceae), *Q. saponaria* (Quillajaceae), *S. maritima* (Surianaceae), *C. pentastylis* (Surianaceae) and *P. persica* (Rosaceae, external group) chloroplasts were used in the DnaSP v6 platform (Rozas et al. 2017) to calculate the nucleotide polymorphisms (Pi) with a window length of 600 bp and a step size of 200 bp. To assess the divergence between chloroplast regions of the five species, we performed a global sequence alignment using the mVISTA platform (Frazer et al., 2004).

## Phylogenetic analysis

Alignment and phylogenetic reconstructions were performed using the function "build" of ETE3 3.1.2 (Huerta-Cepas et al., 2016) as implemented on the GenomeNet (<https://www.genome.jp/tools/ete/>), based on the 13 selected species described in table 1. The sequences were aligned using MAFFT v.7.129 (Katoh and Standley 2013) and the model defined from pmodeltest models using ML tree inference. The tree was generated from RAxML with default parameters, GAMMA JTT/GTR model and 1000 bootstrap replications.

## Results

### *Q. brasiliensis* and *Q. saponaria* assembly

We performed the chloroplast assembly of *Quillaja saponaria* from an Illumina run (Illumina MiSeq) available in accession SRR6297481 with 8,886,340 reads. Our assembly resulted in a chloroplast of 160,528bp in agreement with the assembly by Zhang *et al* (2020) (160,392bp). Assembly and annotation were performed according to the protocol used in *Q. brasiliensis* at GetOrganelle with enough subseven to cover 500x the total value of the assembled genome. As the chloroplast of *Q. saponaria* by Vizoso et al., (2019) was discordant with the size observed in the *Q. brasiliensis* (160360) montage, we chose to use our montage for comparative purposes.

### Overall organization and general features of the plastomes

The species *Q. saponaria* and *Q. brasiliensis* show a pattern in the general size with 160,360 and 160,528 bp, respectively. The Large Single Copy region has 89,482 bp, the SSC region has 19,052 bp and IRs has 25,997 bp (Figure 1). In *Q. saponaria* the LSC region has 89,355 bp and the SSC has 19,022 bp and the IR regions have 25,992 bp. The size and genome organization of the species selected for analysis are described in Table 1. The size of the analyzed chloroplasts ranged from 171,893 bp to 143,284 bp, these values are observed for other species of angiosperms.

We identified 79 genes, 8 rRNAs and 36 tRNAs were found in both species, with the same GC content in both species. The list of genes in the different groups analyzed is described in table 2. For the Surianaceae group we found 80 genes, 8

rRNAs and 54 tRNAs for *C pentastilis* and 82 genes, 8 rRNAs and 56 tRNAs for *S. maritima*, in addition to a GC content of 37 and 35 respectively for these species.

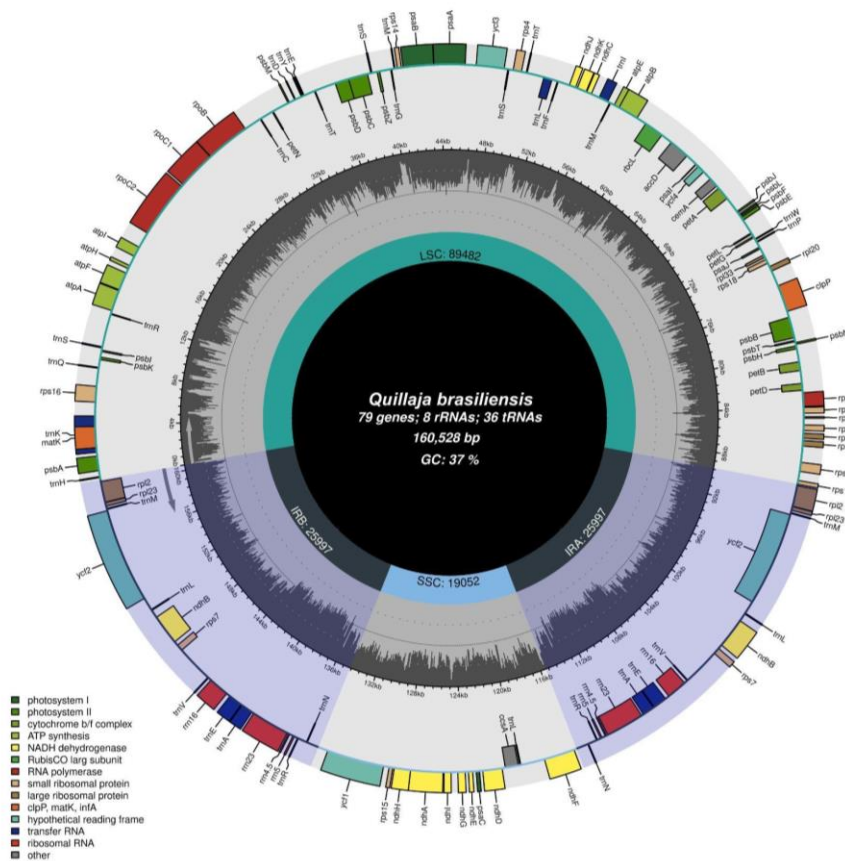


Figure 1: Complete chloroplast of the species *Quillaja brasiliensis*

| Species                        | Length (pb) | LSC      | IRA            | SSC            | IRB           | NCBI accessl |
|--------------------------------|-------------|----------|----------------|----------------|---------------|--------------|
| <i>Quillaja brasiliensis</i>   | 160528      | 1..89482 | 134532..160528 | 115480..134531 | 89483..115479 | this study   |
| <i>Quillaja saponaria</i>      | 160360      | 1..89355 | 134370..160360 | 115347..134369 | 89356..115346 | this study   |
| <i>Cadellia pentastylis</i>    | 157963      | 1..86843 | 132331..157963 | 112477..132330 | 86844..112476 | NC_047334.1  |
| <i>Polygala arillata</i>       | 164947      | 1..83537 | 128348..164947 | 120138..128347 | 83538..120137 | NC_056971.1  |
| <i>Polygala fallax</i>         | 164687      | 1..83352 | 128162..164687 | 119879..128161 | 83353..119878 | NC_052911.1  |
| <i>Polygala japonica</i>       | 165439      | 1..83722 | 128654..165439 | 120509..128653 | 83723..120508 | NC_052912.1  |
| <i>Polygala karensium</i>      | 171893      | 1..73594 | 126949..171893 | 118540..126948 | 73595..118539 | NC_056968.1  |
| <i>Polygala sibirica</i>       | 165397      | 1..83689 | 128603..165397 | 120485..128602 | 83690..120484 | NC_056970.1  |
| <i>Polygala tenuifolia</i>     | 165423      | 1..83699 | 128584..165423 | 120540..128583 | 83700..120539 | NC_050829.1  |
| <i>Salomonina cantoniensis</i> | 143284      | 1..80860 | 116172..143284 | 107974..116171 | 80861..107973 | NC_056969.1  |

| Species                      | Length (pb) | LSC      | IRA            | SSC            | IRB           | NCBI accessi |
|------------------------------|-------------|----------|----------------|----------------|---------------|--------------|
| <i>Suriana maritima</i>      | 164036      | 1..90858 | 137735..164036 | 117161..137734 | 90859..117160 | NC_047313.1  |
| <i>Prunus persica</i>        | 157790      | 1..85882 | 131324..157704 | 112264..131323 | 85883..112263 | NC_014697.1  |
| <i>Epirixanthes elongata</i> | 36275       | *        | *              | *              | *             | NC_046013.1  |

Table 1: Size of the chloroplast regions selected for analysis (\*\*Sequence prepared to enter in the NCBI)(\* undefined regions)

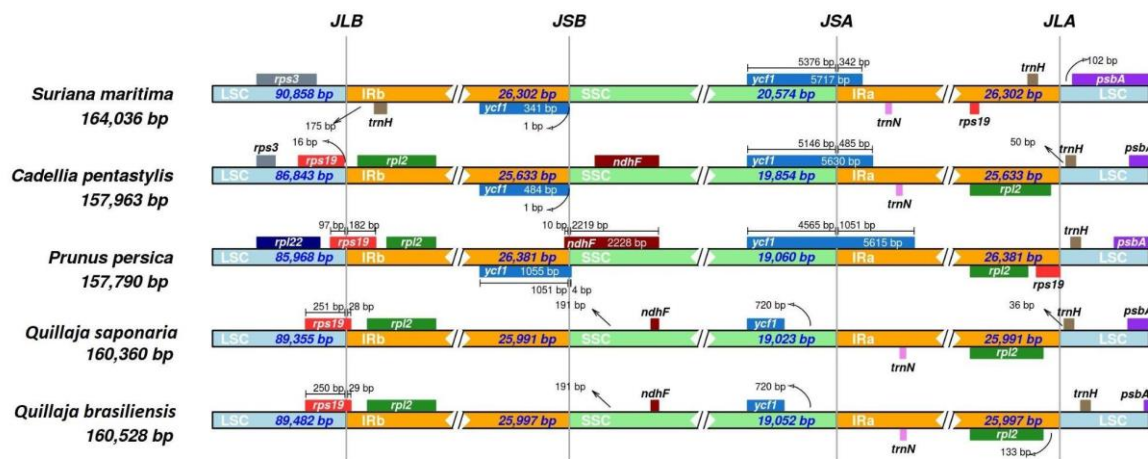


Figure 2: Comparison of the limits LSC, SSC and two IRs of the species *Quillaja brasiliensis*, *Quillaja saponaria*, *Prunus persica*, *Suriana maritima* and *Cadellia pentastylis*. The figure is not to scale in length.

In Polygalaceae, *Polygala falax* does not have the *rpl2* gene, which in turn is duplicated in the other representatives of the family. The *petN* gene is present only in *Poligala falax* and *Poligala arillata*. In surianaceae, there is a duplication of the *rps19* gene in *Suriana maritima* and the absence of the *rps2* and *petB* genes in *Cadellia pentastylis*. The *rps19* gene is duplicated in *Suriana maritima* due to its location in the IR region of this plant. The variation in the number of tRNAs is related to a doubling of the histidine tRNA in *S. maritima*. In Quillajaceae, we found no variation in the content of genes and tRNAs of this species of the family.

We identified the LSC/IRb/SSC/IRa boundaries in the chloroplast genomes of *Q. brasiliensis*, *Q. saponaria*, *Smaritima*, *C. pentastylis* and *P. persica* (Fig. 2). The *ycf1* pseudogene is situated without SSC/IRa limit for the species of Surianaceae, however in Quillajaceae it is entirely located in the SSC region. Except for *S.*

*maritima*, the *rps19* gene has no LSC/IRb limit, in this species we are not lacking the *rpl2* gene.

| Category of genes        |                                    | Families                    |                             |                      |
|--------------------------|------------------------------------|-----------------------------|-----------------------------|----------------------|
|                          |                                    | Quilajaceae                 | Surianaceae                 | Polygalaceae         |
| Genes for photosynthesis |                                    | atpA, atpB, atpE, atpF,     | atpA, atpB, atpE, atpF,     | atpA, atpB, atpE,    |
| Subunits of ATP synthase |                                    | atpH, atpI                  | atpH, atpI                  | atpF, atpH, atpI     |
| Genes for photosynthesis | Subunits of photosystem II         | psbA, psbB, psbC, psbD,     | psbA, psbB, psbC, psbD,     | psbA, psbB, psbC,    |
|                          |                                    | psbE, psbF, psbI, psbJ,     | psbE, psbF, psbI, psbJ,     | psbD, psbE, psbF,    |
|                          |                                    | psbK, psbL, psbM, psbN,     | psbK, psbM, psbN, psbT,     | psbJ, psbK, psbM,    |
|                          |                                    | psbT, psbZ, ycf3            | psbZ, ycf3                  | psbN, psbT, psbZ,    |
|                          |                                    |                             |                             | ycf3                 |
| Genes for photosynthesis | Subunits of NADH-dehydrogenase     | ndhA, ndhB, ndhB,           | ndhA, ndhB, ndhB,           | ndhA, ndhA, ndhB,    |
|                          |                                    | ndhC, ndhD, ndhE,           | ndhC, ndhD, ndhE,           | ndhB, ndhC, ndhD,    |
|                          |                                    | ndhF, ndhG, ndhH, ndhI,     | ndhF, ndhG, ndhH, ndhI,     | ndhE, ndhF, ndhG,    |
|                          |                                    | ndhJ, ndhK                  | ndhJ, ndhK                  | ndhH, ndhH, ndhI,    |
|                          |                                    |                             |                             | ndhI, ndhJ, ndhK     |
| Genes for photosynthesis | Subunits of cytochrome b/f complex | petA, petB, petD, petG,     | petA, petB, petD, petG,     | petA, petB, petD,    |
|                          |                                    | petL, petN                  | petL, petN                  | petD, petG, petL,    |
|                          |                                    |                             |                             | petN                 |
| Genes for photosynthesis | Subunits of photosystem I          | psaA, psaB, psaC, psal,     | psaA, psaB, psaC, psal,     | psaA, psaB, psaC,    |
|                          |                                    | psaJ                        | psaJ                        | psaJ                 |
| Genes for photosynthesis | Subunit of rubisco                 | rbcL                        | rbcL                        | rbcL                 |
| Self replication         | Large subunit of ribosome          | rpl14, rpl16, rpl2, rpl2,   | rpl14, rpl16, rpl2, rpl2,   | rpl14, rpl14, rpl16, |
|                          |                                    | rpl20, rpl23, rpl23, rpl33, | rpl20, rpl23, rpl23, rpl32, | rpl16, rpl2, rpl2,   |
|                          |                                    | rpl36                       | rpl33, rpl36                | rpl20, rpl23, rpl23, |
|                          |                                    |                             |                             | rpl33, rpl36, rpl36  |
| Self replication         | DNA dependent RNA polymerase       | rpoA, rpoB, rpoC1,          | rpoA, rpoB, rpoC1,          | rpoA, rpoA, rpoB,    |
|                          |                                    | rpoC2                       | rpoC2                       | rpoC1, rpoC2         |
|                          |                                    |                             |                             | rps11, rps11,        |
| Self replication         | Small subunit of ribosome          | rps11, rps12, rps12,        | rps11, rps12, rps12,        | rps12, rps12,        |
|                          |                                    | rps14, rps15, rps16,        | rps14, rps15, rps16,        | rps14, rps15,        |
|                          |                                    | rps18, rps19, rps3, rps4,   | rps18, rps19, rps19*,       | rps15, rps18,        |
|                          |                                    | rps7, rps7, rps8            | rps2*, rps3, rps4, rps7,    | rps19, rps19, rps3,  |
|                          |                                    |                             | rps7, rps8                  | rps3, rps4, rps7,    |
|                          |                                    |                             |                             | rps7, rps8, rps8     |
| Other genes              | Subunit of Acetyl-CoA-carboxylase  | accD                        | accD                        | accD                 |
| Other genes              | c-type cytochrom synthesis gene    | ccsA                        | ccsA                        | ccsA                 |
| Other genes              | Envelop membrane protein           | cemA                        | cemA                        | cemA                 |

|                          |                   |                        |                        |      |
|--------------------------|-------------------|------------------------|------------------------|------|
| Other genes              | Protease          | clpP                   | clpP                   |      |
| Other genes              | Maturase          | matK                   | matK                   | matK |
| Unkown<br>reading frames | Conserved<br>open | ycf1, ycf2, ycf2, ycf4 | ycf1, ycf2, ycf2, ycf4 | ycf4 |

Table 2: Gene content in the *sp* genomes of the 13 selected species

| Quilajaceae |       | Surianaceae |       | Poligalaceae |       |
|-------------|-------|-------------|-------|--------------|-------|
| Type        | Codon | Type        | Codon | Type         | Codon |
| Arg         | TCT   | Arg         | TCT   | Arg          | TCT   |
| Cys         | GCA   | Cys         | GCA   | Cys          | GCA   |
| Thr         | GGT   | Thr         | GGT   | Thr          | GGT   |
| Gly         | GCC   | Gly         | GCC   | Gly          | GCC   |
| Ser         | GGA   | Ser         | GGA   | Ser          | GGA   |
| Phe         | GAA   | Phe         | GAA   | Val          | GAC   |
| Met         | CAT   | Met         | CAT   | Arg          | ACG   |
| Val         | GAC   | His*        | GTG   | Leu          | TAG   |
| Arg         | ACG   | Val         | GAC   | Asn          | GTT   |
| Leu         | TAG   | Arg         | ACG   | Leu          | CAA   |
| Asn         | GTT   | Leu         | TAG   | Met          | CAT   |
| Leu         | CAA   | Asn         | GTT   | Gln          | TTG   |
| Met         | CAT   | Leu         | CAA   | Val          | GAC   |
| Val         | GAC   | Met         | CAT   | Arg          | ACG   |
| Arg         | ACG   | His         | GTG   | Asn          | GTT   |
| Asn         | GTT   | Val         | GAC   | Leu          | CAA   |
| Leu         | CAA   | Arg         | ACG   | Met          | CAT   |
| Met         | CAT   | Asn         | GTT   | Gln          | TTG   |
| Pro         | TGG   | Leu         | CAA   | Pro          | TGG   |
| Trp         | CCA   | Met         | CAT   | Trp          | CCA   |
| Thr         | TGT   | Pro         | TGG   | Thr          | TGT   |
| Met         | CAT   | Trp         | CCA   | Met          | CAT   |
| Ser         | TGA   | Thr         | TGT   | Ser          | TGA   |
| Glu         | TTC   | Met         | CAT   | Glu          | TTC   |
| Tyr         | GTA   | Ser         | TGA   | Tyr          | GTA   |
| Asp         | GTC   | Glu         | TTC   | Asp          | GTC   |
| Ser         | GCT   | Tyr         | GTA   | Ser          | GCT   |
| Gln         | TTG   | Asp         | GTC   | Gln          | TTG   |
| His         | GTG   | Ser         | GCT   | Phe          | GAA   |
|             |       | Gln         | TTG   | Met          | CAT   |
|             |       |             |       | His          | GTG   |

Table 3: tRNAs content in the sp genomes of the 13 selected species

### Repeat and SSR analysis

We found 108 repeat regions in *Q. brasiliensis* and 99 in *Q. brasiliensis*. *Quillaja saponaria* showed fewer repetitions in all types with the AAT trinucleotide that remained in two species.

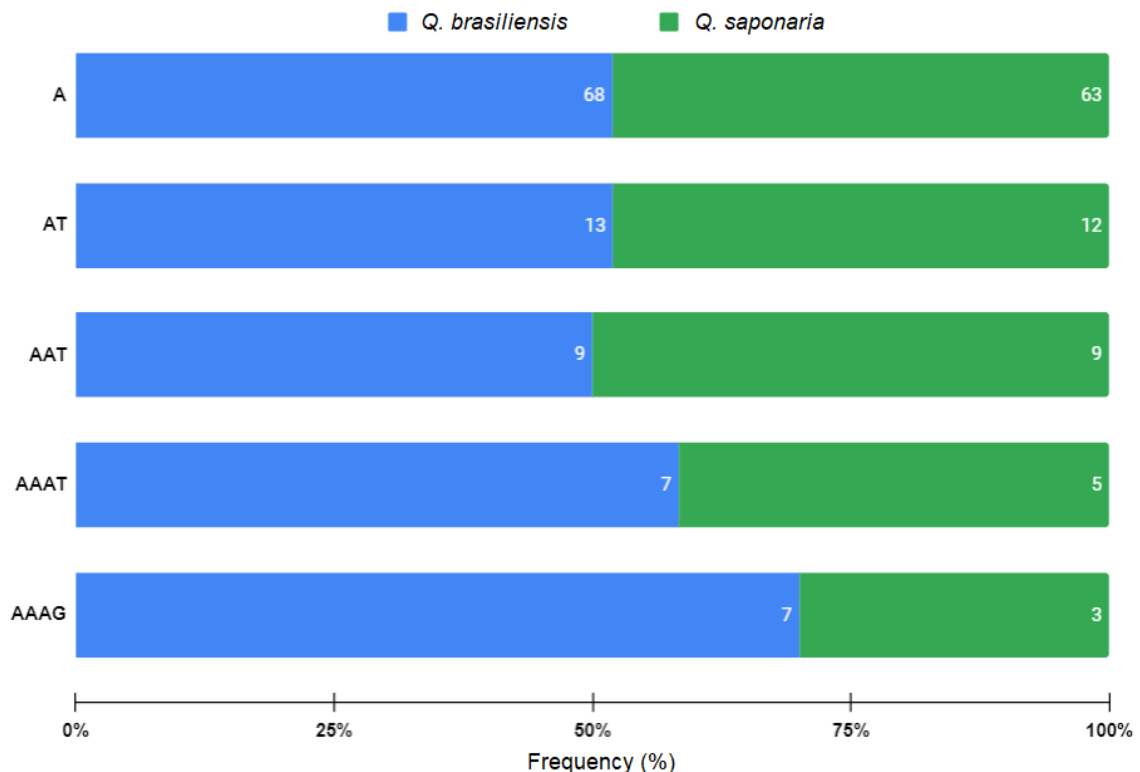


Figure 3: Frequency of SSR in Quillajaceae chloroplast genomes

### Genomic divergence

The IR regions were more conserved than the LSC and SSC regions, and the hypervariables were mainly observed in the single-copy regions. In the selection analysis, the gene groups of the analyzed species were aligned, and the nucleotide substitution rate ( $\omega$ ) evaluated, five of them presented a value of  $\omega$  greater than 1, indicating a positive selection, they are *petB*, *psaI*, *rpl2*, *rpoc1* and *rps7*, shown in figure 5. We identified the level of divergence of the genomic sequences of the chloroplasts of *Q. brasiliensis*, *Q. saponaria*, *S. maritima*, *C. pentastylis* and *P. persica*. We performed a multiple alignment and, using the DnaSP software, calculated the nucleotide variability (Pi). Elevated Pi values ( $>0.08$ ) were identified in the SSC and LSC single copy regions. These high Pi value regions suggested

differences in the comparison between the five cpDNA. The highest value of Pi was identified in the initial region of 5000 to 9000, followed by 130.00 to 132.000 in single copy regions. The most conserved region with the lowest pi value was found in the IR regions. The PI values for the five species are described in figure 4.

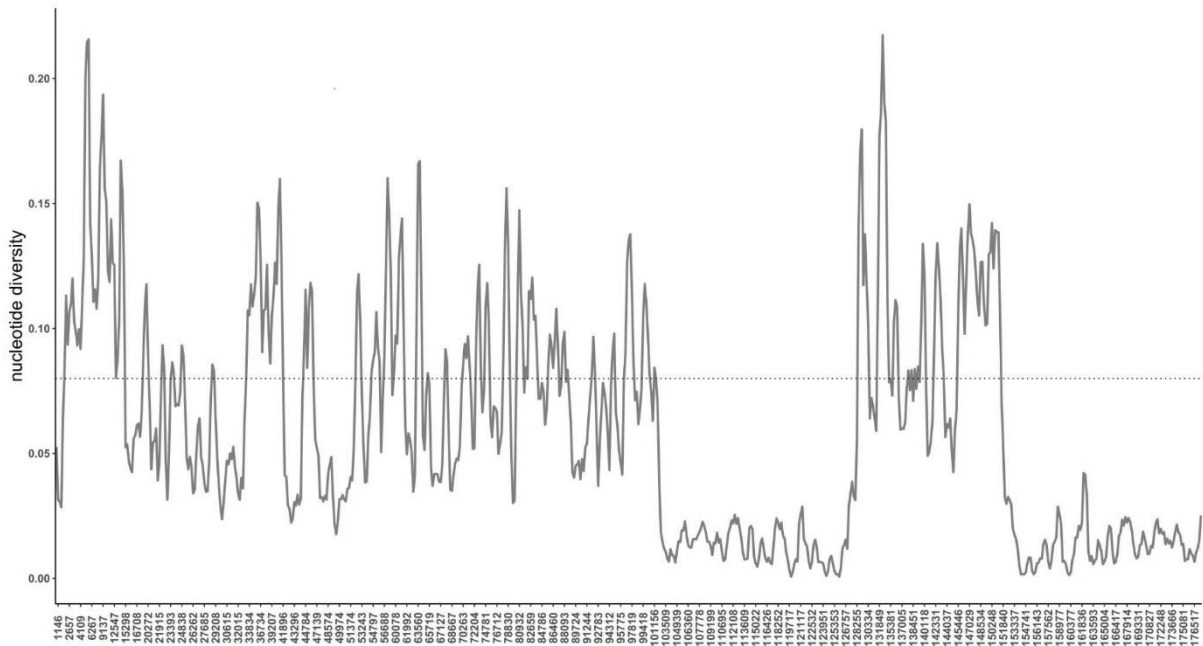


Figure 4: Sliding window analysis of the five *cp* genome of *Q. brasiliensis*, *Q. saponaria*, *Smaritima*, *C. pentastylis* and *P. persica* (window length: 600 bp; step size: 200 bp). X axis: position of the midpoint of a window; Y axis: nucleotide diversity of each window

Figure 5: dN/dS dos genes

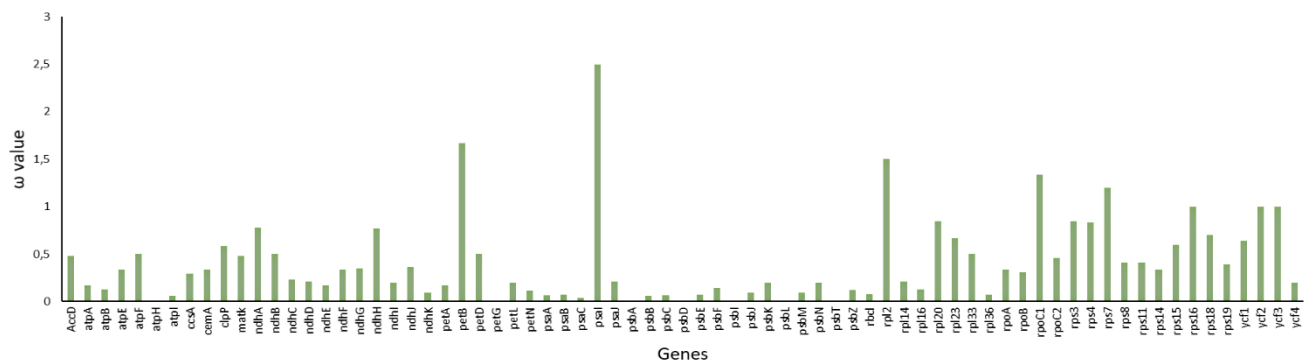


Figure 5: Rate of selection of nucleotides in genes in all species

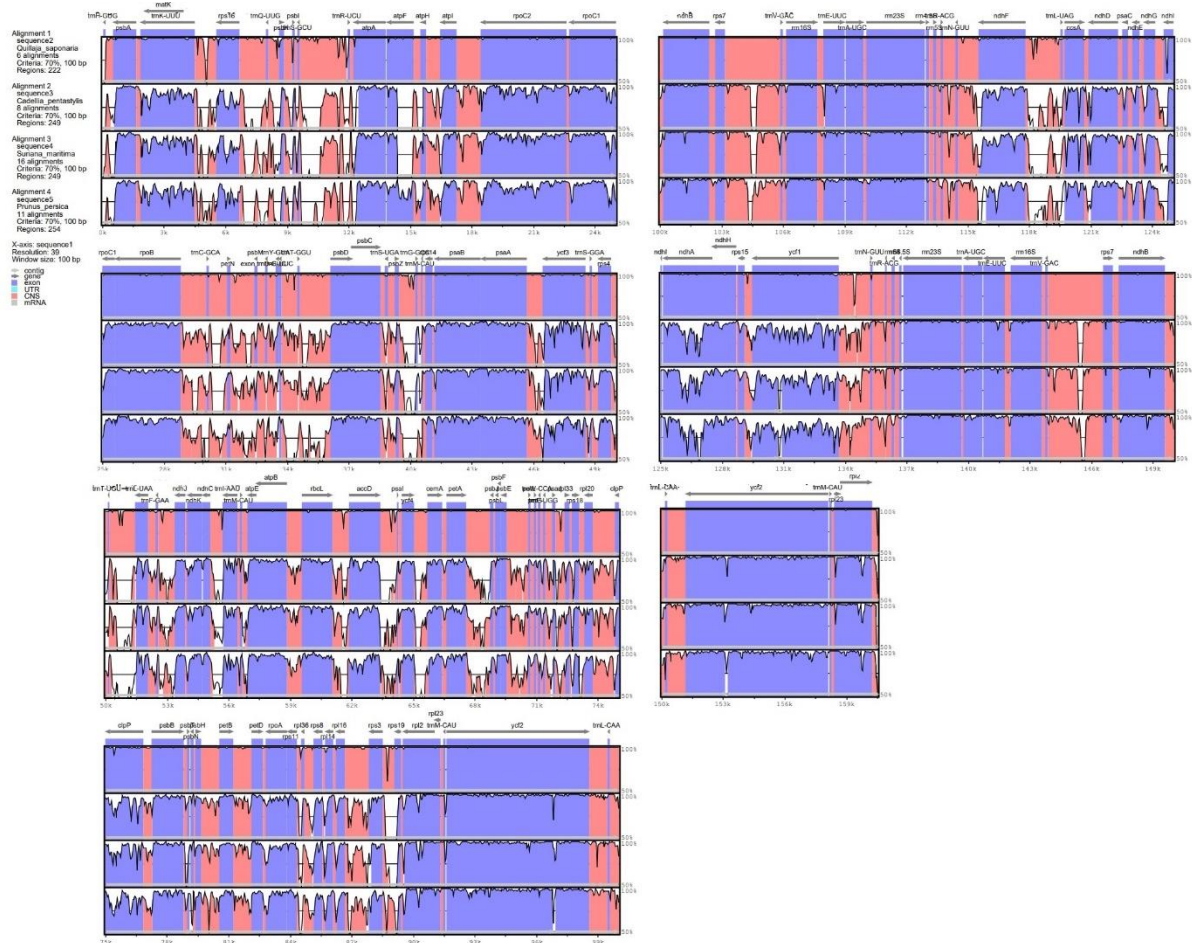


Figure 6: Alignment of the genome sequences of *Q. brasiliensis*, *Q. saponaria*, *Smaritima*, *C. pentastylis* and *P. persica*, using mVISTA. The x-axis represents the baseline sequence of the alignment, and the y-axis represents the percent identity in pairs within 50-100%. Gray arrows represent gene positions and directions

We used the mVISTA platform to perform the alignment of the sequences of the analyzed chloroplasts, and from the annotation of the location of the genes of *Q. brasiliensis* we can identify genic regions and divergent iterencies between the five species (figure 6). The regions that showed the greatest divergence in the comparison of species was the location of the *rps12*, *ycf3*, *ndhA*, *clpP* and *ycf1* genes. Unlike the *ycf1* region, the region where *ycf2* is located in *Q. brasiliensis* presents low divergence in the comparison between the species.

We performed a maximum likelihood (ML) tree with 13 entire plastid genomes to determine the phylogenetic relationships of *Q. saponaria* and *Q. brasiliensis* with related species of Polygalaceae and Surianaceae. We selected 11 complete chloroplast genomes already published from the NCBI *Ephirixanthes elongata* as an outgroup. As shown in the figure, we obtained a phylogeny with two distinct clades, one containing Quillajaceae and Surianaceae and the other containing the species of

Poligalaceae. A representative of rosaceae *Prunus persica* was related to the clade of Surianaceae and Quillajaceae.

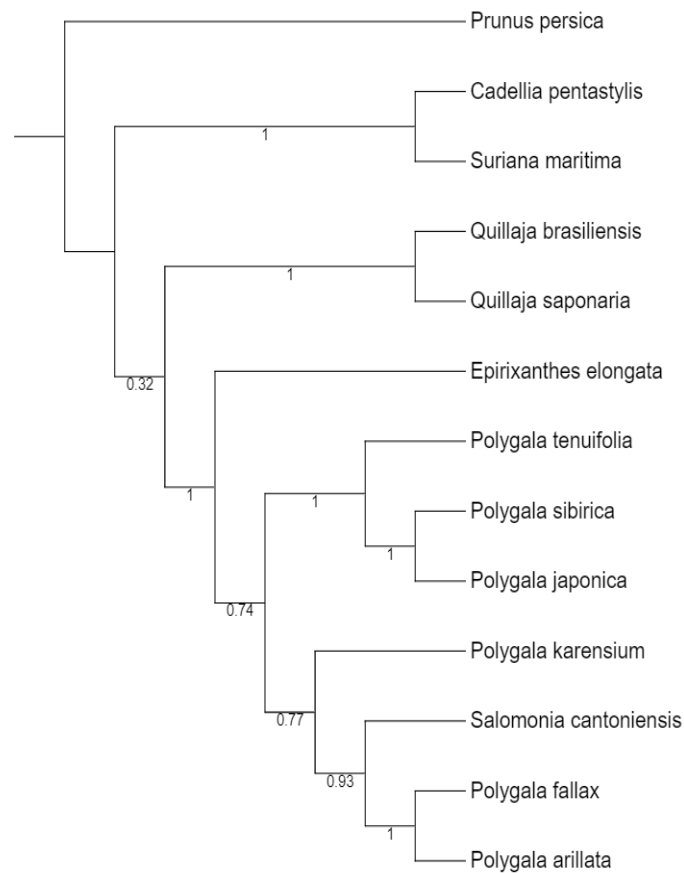


Figure 7: The maximum likelihood (ML) phylogenetic tree with complete chloroplast genome data. Support values are indicated in the branches.

## Discussion

### Changes in the content of plastomes: gene gain/loss

The *atpF* gene does not have introns in *Suriana maritima*, this characteristic is observed only in some representatives of Malpighiales and Rosids (Daniell et al. 2008). In *Quillaja brasiliensis* this gene has introns as in *Q. saponaria*. Regarding size, *Quillaja brasiliensis* complements observations that the plastid genome of *S. maritima* is more closely related to species of Quillajaceae than other Fabales (LAI et al., 2019). The *rpl32* gene is absent in Quillajaceae present only in *Suriana maritima*, *Prunus persica* and *Cadellia pentastylis*. The species *P. Alba* lacks *rpl32*, for this species this gene was reported in the nucleus as a horizontal transfer between plastid genomes (UEDA et al., 2007).

The loss or transfer of the *rpl32* gene in related species may have occurred before the diversification of the Quillajaceae, Surianaceae and Poligalaceae clades. Throughout the evolution of chloroplasts there was a loss of half of the genetic material, this change occurred in relation to environmental changes generating loss of time (XU et al., 2015). Changes in cpDNA content will be more expressive in specific regions, justifying the genetic distribution and size within species. In the case the *ycf2* pseudogene is completely absent in polygala, this gene exceeds 8 kb in IRs region.

### **Variable regions were identified based on plastome sequences**

The *psaI* gene has positive selection indicating a strong selection profile in the genomic structure. This gene is present only in Quillajaceae, Surianaceae and in the external *Prunus persica*. The variation is related to a region of 70 nucleotides more in surianaceae and *Prunus* compared to the *Quillaja* representatives. This gene is normally found upstream of *ycf4*, which is observed in the alignment shown in figure 6, however this gene is absent in some *Lathyrus* species, as well as observed in Polygalaceae, indicating an independent loss between clades (MAGEE et al., 2010).

The *petB* gene is side by side with other *psbT-psbH-petB-petD* genes, they are polycistronic plastid transcripts located in a hypervariable region between 76.500 and 82.000 (MANAVSKI et al., 2018) Being located in a hypervariable region, the gene is under strong selective pressure, in addition to a structural change where the genomic sequences vary from 1423pb in *Prunus persica* and 181pb in *Polygala arilata*. The gene that this positive selection, *rpl2*, has an intron and is at the limit of the IR regions and presents only one intron like other Fabaceae (HONG et al., 2020) It presents positive selection in groups as in *Siraitia* (SHI et al., 2020).

### **Phylogenomic analysis and relationship among species**

The phylogeny obtained from the chloroplast sequences was in agreement regarding the structure and content of the genes observed from the comparison. The clade Surianaceae, Quillajaceae presents the difference of only in the duplication of the *rps19* gene and the presence of the pseudogene *ycf1* and the duplication of the *ycf2* indicating the conservation of the general structure of the plastidia of this clade (DONG et al., 2015).

## Conclusion

The present study carried out the assembly of the chloroplast of the species *Quillaja brasiliensis* and the re-assembly and re-annotation of the chloroplast of *Quillaja saponaria*. With the data obtained, we were able to make the first comparison with all members of the Quillajaceae family with other members of the Fabales order. This study's findings improve our understanding of the evolutionary history of species, with a focus on the plastid genomes in the Quillajaceae family.

## Reference

AYGOREN ULUER, Deniz; FOREST, Félix; HAWKINS, Julie A. Supermatrix analyses and molecular clock rooting of Fabales: exploring the effects of outgroup choice and long branch attraction on topology. **Botany**, v. 98, n. 4, p. 231-247, 2020.

BODYŁ, Andrzej; MACKIEWICZ, Paweł; STILLER, John W. Early steps in plastid evolution: current ideas and controversies. **Bioessays**, v. 31, n. 11, p. 1219-1232, 2009.

CHIESA, Fernando Ferreira et al. Quillaja brasiliensis (A. St.-Hil. & Tul.) Mart. In: **Medicinal and Aromatic Plants of South America Vol. 2**. Springer, Cham, 2021. p. 447-459.

CHOI, In-Su et al. Highly Resolved Papilionoid Legume Phylogeny Based on Plastid. **Insights in Plant Systematics and Evolution: 2021**, 2022.

CUSACK, Brian P.; WOLFE, Kenneth H. When gene marriages don't work out: divorce by subfunctionalization. **Trends in genetics**, v. 23, n. 6, p. 270-272, 2007.

Daniell, H.; Lin, C.-S.; Yu, M.; Chang, W.-J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* 2016, 17, 134.

DA SILVA MAGEDANS, Yve Veronica; PHILLIPS, Michael A.; FETT-NETO, Arthur Germano. Production of plant bioactive triterpenoid saponins: From metabolites to genes and back. **Phytochemistry Reviews**, v. 20, n. 2, p. 461-482, 2021.

DONG, Wenpan et al. ycf1, the most promising plastid DNA barcode of land plants. **Scientific reports**, v. 5, n. 1, p. 1-5, 2015.

HONG, Zhou et al. Comparative analyses of five complete chloroplast genomes from the genus *Pterocarpus* (Fabaceae). **International journal of molecular sciences**, v. 21, n. 11, p. 3758, 2020.

HUERTA-CEPAS, Jaime; SERRA, François; BORK, Peer. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. **Molecular biology and evolution**, v. 33, n. 6, p. 1635-1638, 2016

KUMAR, Sudhir et al. MEGA X: molecular evolutionary genetics analysis across computing platforms. **Molecular biology and evolution**, v. 35, n. 6, p. 1547, 2018.

KUMAR, Arbind et al. Plant-derived immuno-adjuvants in vaccines formulation: a promising avenue for improving vaccines efficacy against SARS-CoV-2 virus. **Pharmacological Reports**, p. 1-17, 2022.

LAI, Qiang et al. Complete plastid genome of *Suriana maritima* L. (Surianaceae) and its implications in phylogenetic reconstruction of Fabales. **Journal of Genetics**, v. 98, n. 5, p. 1-7, 2019.

LEE, Chaehee et al. The chicken or the egg? Plastome evolution and an independent loss of the inverted repeat in papilionoid legumes. **The Plant Journal**, v. 107, n. 3, p. 861-875, 2021.

LEE, Dong-Hyuk et al. The complete chloroplast genome of *Polygala tenuifolia*, a critically endangered species in Korea. **Mitochondrial DNA Part B**, v. 5, n. 2, p. 1919-1920, 2020.

MAGEE, Alan M. et al. Localized hypermutation and associated gene losses in legume chloroplast genomes. **Genome research**, v. 20, n. 12, p. 1700-1710, 2010.

MANAVSKI, Nikolay; SCHMID, Lisa-Marie; MEURER, Jörg. RNA-stabilization factors in chloroplasts of vascular plants. **Essays in biochemistry**, v. 62, n. 1, p. 51-64, 2018.

NEI M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA* 76:5269–5273.

PASTORE, José Floriano Barêa et al. Phylogeny and biogeography of Polygala (Polygalaceae). **Taxon**, v. 68, n. 4, p. 673-691, 2019.

FRAZER, KA, Pachter L, Poliakov A, Rubin E.M, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 2004 Jul 1;32(Web Server issue):W273-9.

ROZAS, Julio et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. **Molecular biology and evolution**, v. 34, n. 12, p. 3299-3302, 2017.

SHI, Hongwu et al. Complete chloroplast genomes of two *Siraitia* Merrill species: Comparative analysis, positive selection and novel molecular marker development. **PloS one**, v. 14, n. 12, p. e0226865, 2019.

SHI, Linchun et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer. **Nucleic acids research**, v. 47, n. W1, p. W65-W73, 2019.

TEMNYKH, Svetlana et al. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome research*, v. 11, n. 8, p. 1441-1452, 2001.

UEDA, Minoru et al. Loss of the *rpl32* gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in *Populus*. **Gene**, v. 402, n. 1-2, p. 51-56, 2007.

VIZOSO, Paula; UNDURRAGA, Soledad Francisca; VELOZO, Juan. Chloroplast genome of the soap bark tree *Quillaja saponaria*. **Frontiers in Ecology and Evolution**, v. 7, p. 104, 2019.

WANG, Pengfei. Natural and synthetic saponins as vaccine adjuvants. **Vaccines**, v. 9, n. 3, p. 222, 2021.

XU, Jian-Hong et al. Dynamics of chloroplast genomes in green plants. **Genomics**, v. 106, n. 4, p. 221-231, 2015.

ZHANG, Rong et al. Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. **Systematic Biology**, v. 69, n. 4, p. 613-622, 2020.

ZHAO, Yiyong et al. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. **Molecular Plant**, v. 14, n. 5, p. 748-773, 2021.

## **Análises transcricionais em *Quillaja brasiliensis*: identificação de genes normalizadores**

### **Introdução**

A Família Quillajaceae pertence à ordem Fabales e é composta apenas por duas espécies: *Quillaja brasiliensis* e *Quillaja saponaria*. A espécie *Quillaja brasiliensis* é nativa do nordeste da Argentina, sul do Brasil e Uruguai. *Q. saponaria* tem sua ocorrência descrita no Chile central (MARTICORENA & MARCHESI 2008, LUEBERT 2013).

No aspecto econômico, esse gênero está relacionado a produção de saponinas. Saponinas costumam ter uma região central hidrofóbica - conhecida como sapogenina - que costuma estar “decorada” com açúcares hidrofílicos. Do ponto de vista biotecnológico, são úteis produção de alimentos como emulsificantes (CHUNG et al., 2017) e atividade de imunoestimulante como adjuvante vacinal (SHARMA et al., 2020; CIBULSKI et al., 2021).

Os extratos vegetais que são utilizados para obtenção das saponinas são oriundos da casca de *Q. saponaria* por método destrutivo (BUSTIN et al., 2017). Estudos recentes com *Quillaja brasiliensis* destacaram a presença de saponinas em maior quantidade na folha (YENDO et al., 2015) além de característica imunoestimulante comparável a *Q. saponaria* (CIBULSKI et al., 2018; CIBULSKI et al., 2016). No entanto, são escassos os dados transcricionais da espécie *Q. brasiliensis*, assim como dados genômicos das vias de biossíntese de saponinas para essa espécie.

A técnica de RT-qPCR é a forma mais eficaz de quantificação da expressão de genes em amostras de RNA (NAVARRO et al., 2015), no entanto para que essa quantificação seja realizada é preciso alguns passos anteriores como a elucidação do transcriptoma ou genoma do organismo e a utilização de genes normalizadores desta análise (JOSEPH et al., 2018). Para a validação da expressão de transcritos em RT-qPCR ou identificação de genes chaves para vias metabólicas de interesse, é preciso a identificação de genes considerados normalizadores, ou seja, genes que independente da condição experimental apresentam o mesmo padrão transcricional

(BUSTIN et al., 2017). Normalizadores possibilitam a reprodutibilidade dos experimentos que utilizam das técnicas de RT-qPCR e a supressão de possíveis interferências nas análises de expressão gênica (ZHOU et al., 2017).

O objetivo deste estudo foi selecionar genes candidatos a normalizadores de RT-qPCR em *Quillaja brasiliensis*, com base em dados de um transcriptoma de referência, e avaliar sua estabilidade transcricional em folhas, caule e raiz.

## **Material e métodos**

### **Material vegetal e extração de RNA**

Sementes foram coletadas de árvores adultas de *Quillaja brasiliensis* em Canguçu, RS, Brasil (31°23'42"S, 52°40'32"W). Elas foram germinadas e mantidas sob lâmpadas fluorescentes brancas a  $25 \pm 2$  °C em um ciclo de 16/8 h dia/noite, no Laboratório de Fisiologia Vegetal (Departamento de Botânica, Instituto de Biociências, Universidade Federal do Rio Grande do Sul). Órgãos e tecidos vegetativos foram coletados deste material a partir de 11 meses após o início da germinação. O espécime *voucher* está depositado no Herbário Universitário da UFRGS (ICN 142953).

Foram realizados dois experimentos para extração de RNA total. O Experimento I foi realizado para obtenção de RNA total para sequenciamento do transcriptoma e o Experimento II foi realizado para obtenção de RNA para avaliação da estabilidade de genes normalizadores.

No experimento I, foram extraídas amostras de RNA de folhas submetidas ao estresse por UV-C em tratamento baseado no realizado por Costa et al. (2013) e culturas celulares. Em resumo, o tratamento com UV-C foi realizado em folhas maduras, que foram submetidas ao tratamento por luz UV-C durante 15 minutos. As amostras controle estiverem pelo mesmo período sob luz branca fluorescente. Para as culturas celulares, o ciclo inicial teve início em 17/11/2020. A partir dessa data inicial, foram obtidas amostras de culturas celulares em três fases: Fase LAG: quatro dias após o cultivo inicial; Fase LOG: sete dias após o cultivo inicial; Fase ESTACIONÁRIA: 20 dias após o cultivo inicial. Folhas e culturas celulares foram coletadas e tiveram RNA total extraído em triplicata. No Experimento II foram obtidos RNAs totais de folhas maduras, caule e raiz, todos em condições fisiológicas usuais e sem manifestação de doenças.

A extração de RNA total foi realizada de acordo com o protocolo de WHITE et al., (2008). O material foi tratado para degradação de DNA genômico residual com a DNase TURBO (Thermo). Para avaliação da integridade do RNA extraído, as amostras foram checadas por eletroforeses em gel de agarose. A quantificação do RNA foi realizada por fluorimetria (Qubit, Life Technologies).

Estes passos foram realizados em colaboração com a equipe do prof. Dr. Arthur Germano Fett Neto (Centro de Biotecnologia, UFRGS).

### **RNA-seq: sequenciamento Iso-seq e Illumina**

Para a técnica de sequenciamento PacBio Iso-seq, o RNA total de todas as amostras do Experimento I foram reunidas em um pool equimolar. O material foi recebido pela empresa Novogene (Pequim, RPC), a qual avaliou a qualidade da amostra por métodos fluorimétricos e espectrofotométricos.

O RNA total do pool equimolar foi então utilizado para construção de bibliotecas Iso-seq. A primeira fita de DNA complementar (cDNA) foi sintetizada usando o kit de síntese de cDNA ClontechSMARTer PCR (ClonTech, Takara Bio Inc., Japão) com oligo(dT) como primer. A segunda fita do cDNA foi gerada por PCR em larga escala otimizado. O cDNA de fita dupla foi utilizado para elaboração de bibliotecas SMRTbell com o kit PacBio SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, EUA), seguindo os protocolos do fabricante. Uma biblioteca SMRTbell, sem seleção prévia de tamanho de transcritos, foi sequenciada na plataforma PacBio Sequel III. Todos estes procedimentos foram realizados pela empresa Novogene (Tianjin, China).

Para o RNA-seq Illumina, foram sequenciadas 15 bibliotecas: 5 condições (folhas em condições controle, folhas tratamento UV-C, culturas celulares em fase LAG, culturas celulares em fase LOG, culturas celulares em fase estacionária) em triplicatas biológicas.

Para a construção de bibliotecas Illumina foram novamente avaliados, para cada biblioteca, padrões de integridade do RNA a ser sequenciado (RIN acima de 7,0). A partir do RNA total de cada amostra, foi isolado RNA com cauda poli-A, com o uso de beads magnéticos com esferas oligo(dT). As amostras de RNA poli-A foram então usadas para a construção de bibliotecas RNA-seq usando o NEB Next Ultra RNA Library Prep Kit for Illumina (New England Biolabs), de acordo com as instruções do fabricante. As 15 bibliotecas de RNA-seq foram sequenciadas (150 pb,

paired-end) em uma plataforma Illumina NovaSeq 6000 em Davis, EUA, pela empresa Novogene.

### **Montagem dos transcritos e quantificação de expressão**

A partir dos dados obtidos de Iso-seq, foi executada uma abordagem híbrida para definição de um transcriptoma de referência. Ela foi baseada no *pipeline* IsoSeq v3 (<https://github.com/PacificBiosciences/IsoSeq/blob/master/isoseq-clustering.md>), associada a correção dos transcritos por meio de uma abordagem que utiliza as sequências Illumina obtidas separadamente por tecido, por meio do programa LoRDEC (Salmela e Rivals, 2014). Para a definição final de transcritos não redundantes, foram utilizadas as ferramentas Cogent e cDNA CupCake TOFU (disponíveis em <https://github.com/Magdoll/>).

Para inferência da completude do transcriptoma obtido, foi utilizada a ferramenta BUSCO v5 (do inglês, *benchmarking universal single-copy orthologs*) (MANNI et al., 2021), utilizando as embriófitas como grupo de referência (embryophyta\_odb10.2020-09-10).

Para detecção de sequências codificadoras (CDS; do inglês *coding sequences*), as sequências foram submetidas à predição do *software* CodAn (NACHTIGALL et al., 2021). A coleção de scripts desenvolvida para a realização das etapas acima foi disponibilizada em um positório público no GitHub (<https://github.com/alanlorenzetti/isoquill>).

Para avaliação do nível transcricional dos genes obtidos no transcriptoma, foi utilizado o *software* Kallisto (BRAY et al., 2016), tendo como input as leituras de RNA-seq Illumina contra o transcriptoma não redundante de referência obtido por meio do método Iso-Seq (PacBio). Kallisto é um método de quantificação de RNA-Seq “livre de alinhamento”. Os resultados do programa foram normalizados para contagens.

As contagens normalizadas para cada gene (TPM; do inglês *transcripts per million*), foram colapsadas entre as réplicas biológicas com o intuito de obter um único valor médio para cada condição ( $\bar{x}_{\text{TPM}}$ ). Utilizamos estes valores de TPM para medir os níveis de expressão para cada gene em todas as amostras. O objetivo, neste caso, foi determinar genes candidatos para uso como normalizadores.

Quatro critérios foram utilizados para a seleção de genes de normalizadores: o gene deveria ser expresso em todos os tecidos avaliados, com baixa variação dos

valores de TPM considerando o desvio padrão ( $[\log_2(\text{TPM})] < 1$ ), nível de expressão médio a alto ( $[\log_2(\text{TPM})] > 5$ ) e sem variação significativa de expressão em termos estatísticos ( $q\text{-valor} > 0,1$ , no pacote DESeq). Estas indicações foram utilizadas por Li et al., (2019) e JURECKOVA et al., (2021), para a definição de genes normalizadores com base em dados de RNA-seq. Genes que correspondiam a estas condições foram anotados na plataforma TRAPID 2.0 (Bucchini et al., 2021).

Esta etapa do trabalho foi realizada pelo Dr. Alan Rodrigues Lorenzetti, técnico em Bioinformática do Laboratório quando da realização deste estudo.

### **Seleção de genes normalizadores e desenho de *primers***

Foram selecionados manualmente, com base na função biológica do gene e menor coeficiente de variação de expressão (item 2.2), um total de 13 genes para desenho de primers e teste de amplificação.

Após a seleção de genes (indicada no item 3.2.1 de resultados), os iniciadores foram desenhados com o software Primer 3 (UNTERGASSER et al., 2012), nos quais os amplicons foram calculados para ter entre 100 e 200 pb, e temperatura de anelamento dos iniciadores entre 52°C e 63°C, com ótimo em 60°C.

Antes do emprego dos genes normalizadores em RT-qPCR, o padrão de amplificação dos genes candidatos a normalizadores foi avaliado por PCR convencional.

### **Síntese de cDNA e PCR convencional**

O RNA total do experimento II foi utilizado para a síntese de cDNA com o kit GoScript™ Reverse Transcription System (PROMEGA), de acordo com as instruções do fabricante, utilizando-se 4 µg de RNA total nas reações. Para os testes de amplificação foi utilizado cDNA de folha. A amplificação por PCR convencional foi realizada com o kit GoTaq® G2 Flexi DNA Polymerase (Promega), utilizando-se 1,0mM de cloreto de magnésio, 0,2mM de dNTPs, 1,0µM de cada primer e 125 ng de cDNA.

As amostras foram amplificadas por PCR convencional no termociclador Veriti (Thermo Fisher), com as seguintes condições de amplificação: uma desnaturação inicial de 2min a 94°C, 40 ciclos a 95°C por 10s, 55°C por 30s e 72°C por 30

segundos, seguidos por uma amplificação final de 72°C por 5min e manutenção final a 4°C.

Os resultados obtidos foram avaliados qualitativamente por eletroforese em gel de agarose (1,5%), corrido em tampão TBE 1X, por 120 min a 90 volts. Foi utilizado no gel de agarose 5µL de uma reação de PCR de 20µL para avaliação da amplificação. Utilizou-se o padrão de massa molecular de DNA de 50bp (Ladder Plus, Sinapse Biotecnologia) para avaliação do tamanho dos *amplicons* obtidos.

Com base nos resultados de amplificação, três genes foram selecionados para avaliação como normalizadores por RT-qPCR.

### RT-qPCR

Com base nos resultados de amplificação por PCR convencional, foram selecionados três genes para as análises de RT-qPCR: Heat shock protein Hsp90 (HSP), Ubiquitina (UBQ) e Ciclofilina (CYP), detalhados na seção 2.2.1 de Resultados e Discussão. Foram utilizados cDNAs do Experimento II para análise por qRT-PCR.

As reações de RT-qPCR foram realizadas no sistema QuantiStudio3 (Applied Biosystems). Para cada reação de qPCR, utilizamos 8,5 µL de qPCRmix (Promega, Madison, WI, EUA), 1 µL de cada iniciador (1 µM), 3,5 µL de cDNA (10 ng / µL) e 0,5 µL de CXR em um volume final de 12 µL. Uma reação sem qualquer cDNA foi incluída como controle negativo para cada par de primers. As condições de RT-qPCR foram de: ciclo inicial de 95 ° C por 2 min; seguido de 40 ciclos a 94 ° C por 30s e 60 ° C por 60s. As curvas de fusão de *amplicons* foram avaliadas para verificar a presença de um único produto com a presença de um controle negativo para cada primer testado.

As reações foram realizadas com três réplicas biológicas e três replicatas técnicas, seguindo as diretrizes do MIQE para experimentos de RT-qPCR (TAYLOR et al., 2010).

### 2.7 Estabilidade de genes de referência

Utilizando-se os dados de CT obtidos na análise de qRT-PCR, foi avaliada a estabilidade média de expressão dos genes normalizados na plataforma RefFinder (<http://www.heartcure.com.au/reffinder/>). O RefFinder é uma ferramenta Web

desenvolvida para avaliar genes de referência em conjuntos de dados experimentais. Ele integra os principais programas computacionais atualmente disponíveis: geNorm (VANDESOMPELE et al., 2002), Normfinder (ANDERSON et al., 2004), BestKeeper (PFAFFL et al., 2004) e o método comparativo Delta-Ct (SILVER et al., 2006) para comparar e classificar os genes de referência candidatos testados. Foram ainda avaliadas as eficiências de amplificação dos iniciadores testados no programa LinReg (RUIJTER et al., 2009).

## Resultados e discussão

### Sequenciamento Iso-seq: estatísticas gerais

O sequenciamento de RNA utilizando a tecnologia Iso-Seq visou gerar um transcriptoma completo e de alta qualidade para *Quillaja brasiliensis*, que servisse de ponto de partida para as análises subsequentes. O sequenciamento Iso-seq gerou aproximadamente 4,4 milhões de Zero Mode Waveguides (ZMWs), totalizando aproximadamente 80 milhões de *subreads*. Em total de bases sequenciadas, as *subreads* totalizaram aproximadamente 150 bilhões de bases.

O tamanho médio das *subreads* foi de 1.894 nt e o N50 de 1.948 nt, indicando que 50% das bases sequenciadas estão contidas em transcritos de tamanho igual ou maior que esse. Um histograma de tamanho das *subreads* é apresentado na

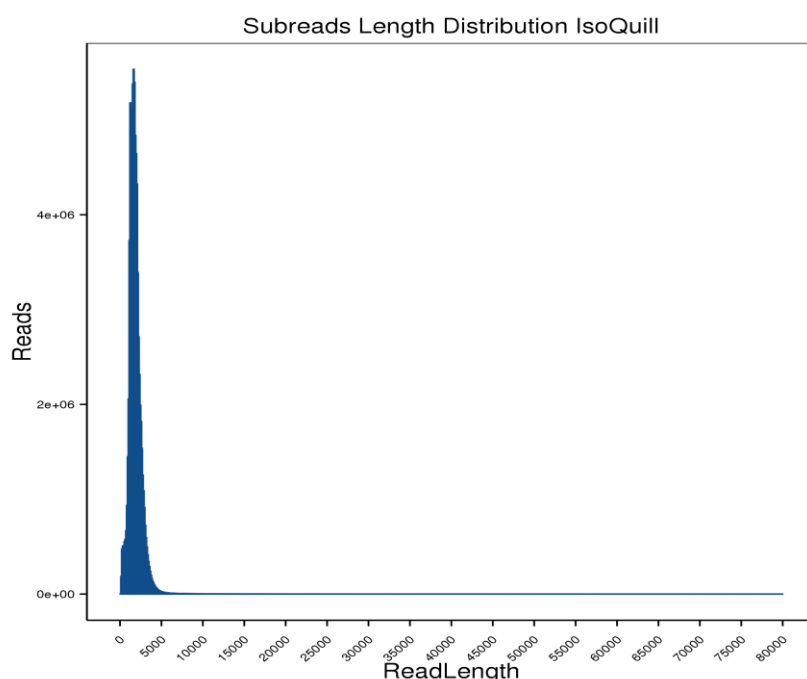


Figura 1: Tamanho das *subreads* sequenciadas. No eixo-y, a frequência absoluta das *reads*; no eixo-x, cada classe de tamanho. Fonte: Novogene (Pequim, RPC).

## Figura 1

Na etapa de obtenção das sequências consenso circulares (CCS; do inglês, *circular consensus sequences*), foram eliminadas 56% das ZMWs, às quais apresentaram conjuntos de *subreads* que não satisfizeram os cortes de número de passagens e acurácia predita. Na etapa seguinte, com aproximadamente 2 milhões de ZMWs, 15% foram removidas por apresentarem primers não identificáveis ou em ordem inesperada. Das 1,6 milhão de ZMWs restantes, 20 mil ainda foram descartadas por serem quiméricas ou desprovidas de cauda poli-A. Para a etapa de *clustering* dos transcritos *full-length non-chimeric* (do inglês, FLNC), foram utilizados 1.610.026 *reads*, a partir dos quais foram obtidos 162.783 clusters de alta qualidade. Esses clusters foram processados para a remoção de redundância utilizando os softwares Cogent e Cupcake-ToFU (disponíveis em <https://github.com/Magdoll/>) e obtivemos 56.431 sequências não redundantes, que representam o transcriptoma-referência de *Quillaja brasiliensis* utilizado neste estudo. Um sumário dos resultados das etapas descritas é apresentado na Tabela 1.

Tabela 1: Sumário do número de *reads* e algumas propriedades do transcriptoma não redundante de *Q. brasiliensis*.

| Parâmetro                   | Frequência absoluta |
|-----------------------------|---------------------|
| <i>Subreads</i> (raw reads) | 79.076.758          |
| ZMWs                        | 4.385.393           |
| CCS reads                   | 2.458.927           |
| FLNCs                       | 1.610.026           |
| Clusters de transcritos     | 162.783             |
| Transcritos colapsados      | 56.431              |
| Média de tamanho (nt)       | 1.961               |
| Sequência mais curta (nt)   | 102                 |
| Sequência mais longa (nt)   | 6.621               |
| Extensão N50 (nt)           | 2.126               |

Subsequentemente, empregamos a abordagem BUSCO (MANNI et al., 2021) e 78% (1.258/1.614) dos ortólogos universais de cópia única das embriófitas estão representados nessas sequências (Tabela 2). Isso indica uma boa completude do transcriptoma, considerando a quantidade de condições avaliadas (apenas dois tecidos). Ressaltamos ainda a grande quantidade de elementos completos duplicados, que é maior do que a quantidade de elementos completos de cópia única, o que se explica pela presença de diferentes isoformas de transcritos para um mesmo gene.

Tabela 2: Resultados da análise de completude. a Frequência absoluta. b Frequência

| Termo BUSCO                 | <sup>a</sup> Freq. absoluta | <sup>b</sup> Freq. relativa |
|-----------------------------|-----------------------------|-----------------------------|
| Completo (C)                | 1.258                       | 78%                         |
| Completo de cópia única (S) | 368                         | 22,9%                       |
| Completo duplicado (D)      | 889                         | 55,1%                       |
| Fragmentado (F)             | 48                          | 3%                          |
| Não achados (M)             | 308                         | 19%                         |

relativa ao total de 1.614 ortólogos universais de cópia única de 50 genomas de embriófitas.

Realizamos a predição de sequências codificadoras (CDS) no conjunto de transcritos não-redundantes. Do total de 56.431 transcritos, identificamos que 52.177

(93%) possuem CDS completas e 1003 (2%) possuem CDS parciais. Os 2.948 transcritos (5%) remanescentes podem ser considerados potenciais RNAs não codificadores (ncRNAs; do inglês, non-coding RNAs).

### **Normalizadores e análises transcricionais**

#### **Identificação de candidatos a genes normalizadores para RT-qPCR**

Utilizando a estimativa de abundância de expressão para cada gene feita com dados de RNA-seq do sequenciamento Illumina (item 2.2 de Material e métodos), foi realizada a predição de candidatos a genes de referência para quantificação por RT-qPCR (item 2.6 de Material e Métodos).

Foram identificados previamente 588 genes compatíveis com as características desejáveis para um normalizador, segundo os critérios do estudo de LI et al., (2019). A partir destes 588 genes, foi possível selecionar um conjunto de genes mais apropriados para normalização, tomando-se por referência os com menor coeficiente de variação, bem como os com função biológica normalmente presente nos genes normalizadores.

Com base nesses referenciais, 13 genes, detalhados na Tabela 3, foram selecionados para o desenho de iniciadores. Realizamos uma avaliação da amplificação dos primers para estes genes em PCR convencional, cujos resultados são apresentados na Figura 2. Baseado na função gênica usual de genes normalizadores (CHEN et al., 2018; DUAN et al., 2020; SONI et al., 2021), bem como o padrão de banda única nas amplificações por PCR, três genes foram selecionados para os testes de RT-qPCR: 1) Heat shock protein Hsp90 (*HSP*), 2) Cyclophilin-like domain (*CYP*) e 3) Ubiquitin (*UBQ*).

Tabela 3: Genes candidatos a normalizadores com respectivos pares de primers

| Símbolo | Nome do gene                                 | Sequência do primer (5' - 3')                               | ID gene  | Tamanho do amplicon |
|---------|----------------------------------------------|-------------------------------------------------------------|----------|---------------------|
| UBQ     | Ubiquitin                                    | F - GCCCATGCCGTCGTGAATC<br>R - AAATGCAACGACGATGACACCA       | PB.30    | 153                 |
| HST72   | Histone-fold H2B                             | F - TCAGGGTGGACTTGCTTCAGAA<br>R - AAATCGACGGTTGCGGAAAAGG    | PB.4066  | 173                 |
| 60S     | 60S acidic ribosomal protein                 | F - GCAACCTCTTCACTGACCTCCT<br>R - TTTGCGTGGTGAAGTCAAGTGGTA  | PB.14445 | 177                 |
| CYP     | Cyclophilin-like domain                      | F - ATCGATTCTCCGCCTGTTCCAT<br>R - GAATTTCCGGGCTCTTTCGACC    | PB.1999  | 153                 |
| HSP90   | Heat shock protein Hsp90                     | F - AGTAATCTTGGTACCCCTGCCA<br>R - TGGTCAGTTTGGTGTGTTGGGTTT  | PB.9553  | 172                 |
| RBSL    | Ribosomal protein L23/L15e core domain       | F - CTTTCCCTTTGCCGCATCTGAG<br>R - ATGCCAAGAAGTACGAGCCCAA    | PB.2856  | 155                 |
| MEF2    | Transcription factor MEF2-like               | F - GGGCCATTGGAATTGCAGACATT<br>R - GCAACTGATGGAAGAGAATGAGCG | PB.17572 | 150                 |
| RLA     | Rossmann-like alpha/beta/alpha sandwich fold | F - CAGCGTCGAAGGAGGAGTAGAT<br>R - GAGGAAGATTGTGGTGGCGGTG    | PB.14498 | 152                 |
| SH3     | Translation protein SH3-like domain          | F - ATGTCCATGCTTGCCAGTTTTG<br>R - GAGCACCCTTCGAGTCCAAAG     | PB.4054  | 153                 |
| EIF-1A  | Translation initiation factor 1A (eIF-1A)    | F - TTCAGAAGCCTAGCCTCATCGG<br>R - AATGGTCGATGTGAGGCTACGT    | PB.19742 | 182                 |
| RBS27   | Ribosomal protein S27e                       | F - AATTGCCGCACACCACAACAG<br>R - GCTTAACCCACCAGCTGAGC       | PB.261   | 155                 |
| RBSS    | Ribosomal protein S7 domain                  | F - TCTGCCATTCTATGAGTCTCTTCC<br>R - CGTCCGGGTGCGAAATATGGCTT | PB.12671 | 107                 |
| SUM     | Ubiquitin domain (SUMO-like domain)          | F - ACGGCGACCATCAAACAAGAAG<br>R - AGAGGAGGACAAGAAGCCCAAC    | PB.23567 | 184                 |

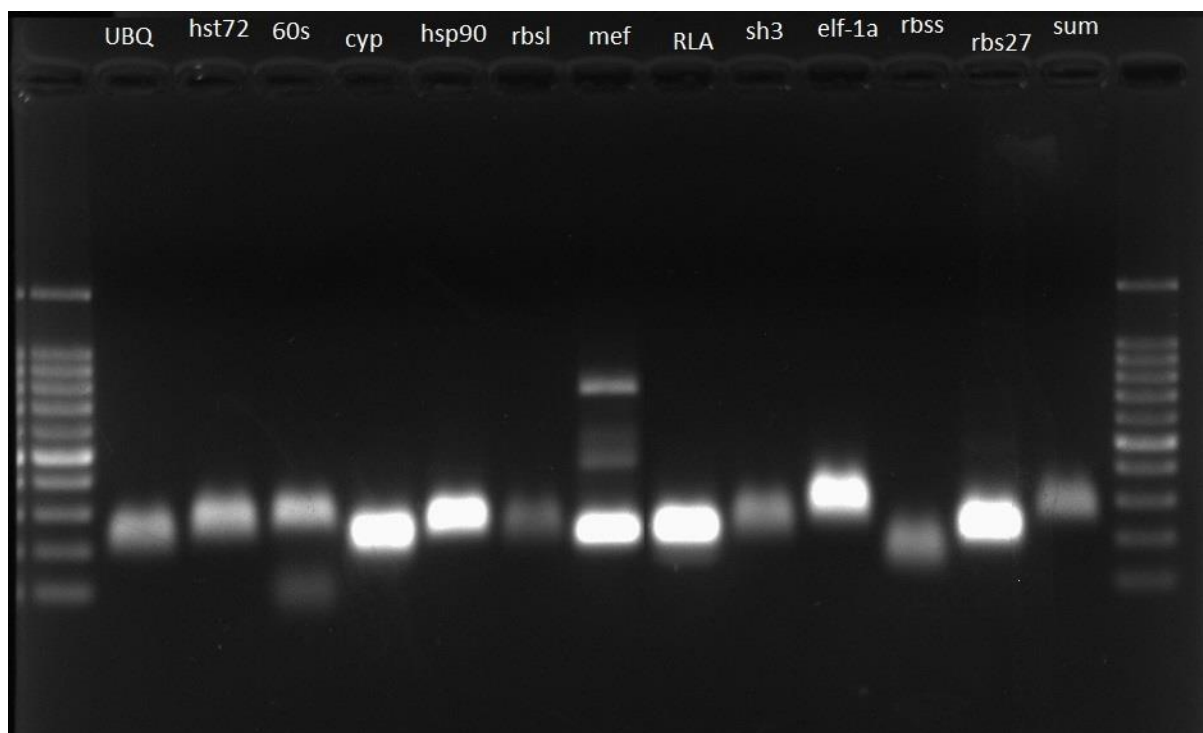


Figura 2: Perfil de amplificação de genes em PCR convencional. Os métodos de amplificação e o material utilizados estão detalhados no item 2.5 de Materiais e Métodos. À esquerda e à direita, é utilizado o padrão de 50pb (50Bp Ladder Plus, sinapse biotecnologia)

### Expressão relativa e estabilidade de normalizadores

O método de quantificação de genes por RT-qPCR apresenta alta sensibilidade, no entanto é preciso de genes de referência como normalização dos dados (BUSTIN et al., 2017). O gene selecionado como referência deve apresentar estabilidade entre diferentes condições e tecidos, assim o teste com diferentes normalizadores possibilita a seleção mais adequada para os diferentes tipos de amostragem que utilizam o método de RT-qPCR (JOSEPH et al., 2018).

Os genes normalizadores selecionados com base nos dados de RNA-seq e padrão de amplificação por PCR foram avaliados quanto a seu padrão transcricional e estabilidade em RNA de caule, folhas maduras e raízes secundárias. Os maiores níveis de expressão, denotados pelo menor CT médio, foram observados em folhas.

O gene CYP em folha foi o que apresentou menor valor de CT ( $20,36 \pm 6,58$ ); o maior valor de CT foi observado no gene UBQ em folha ( $34,11 \pm 4,8$ ). Os demais valores de expressão relativa (média  $\pm$  erro-padrão) estão presentes na Tabela 4 e 3.

| Tecido | Gene | Média CT e Erro |
|--------|------|-----------------|
| Folha  | UBQ  | 21,74 ± 3,53    |
|        | CYP  | 20,36 ± 6,58    |
|        | HSP  | 21,44 ± 6,77    |
| Caule  | UBQ  | 34,11 ± 4,8     |
|        | CYP  | 21,59 ± 6,52    |
|        | HSP  | 24,7 ± 7,22     |
| Raiz   | UBQ  | 22,69 ± 5,57    |
|        | CYP  | 25,97 ± 4,11    |
|        | HSP  | 22,47 ± 5,57    |

Tabela 4: Média e erro padrão dos valores de expressão dos genes normalizadores avaliados por RT-qPCR.

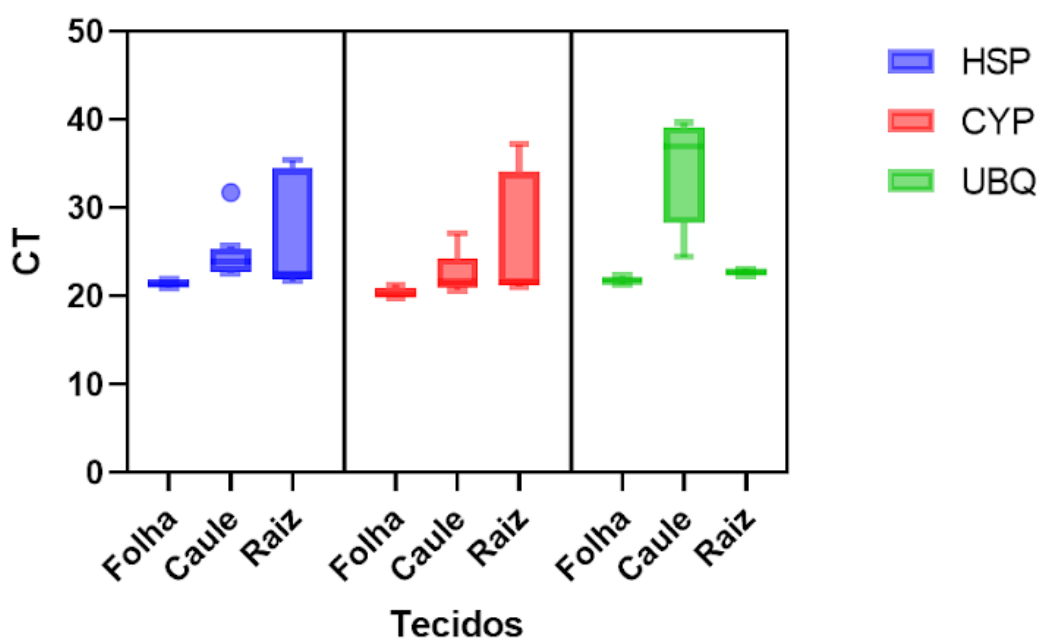
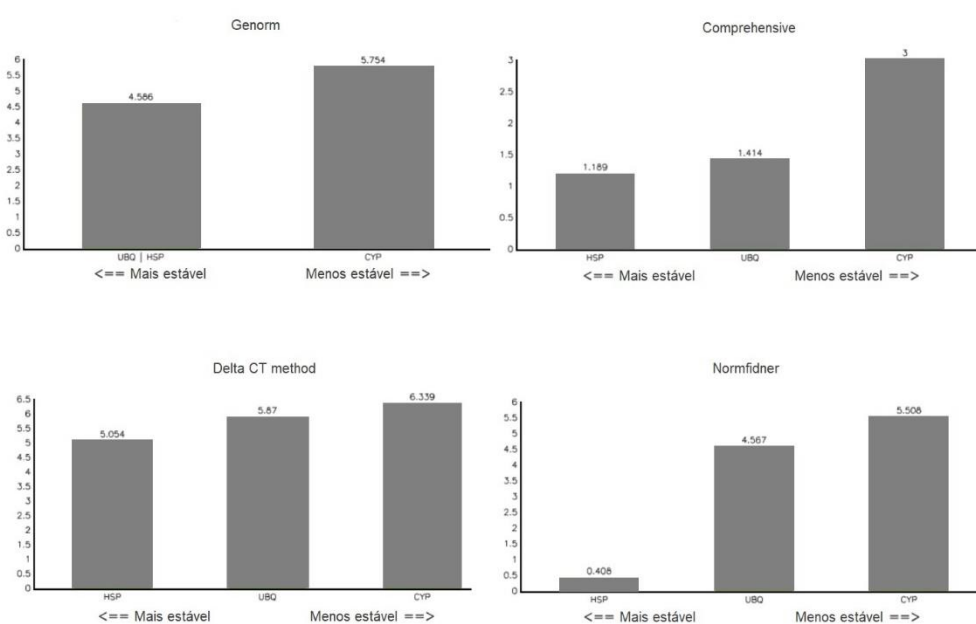


Figura 3: CT dos genes avaliados por RT-qPCR em *Q. brasiliensis*, por tecido. O gráfico Boxplot mostra os valores médios como linhas nas caixas e as barras mostram os valores máximos e mínimos.

A avaliação de estabilidade realizada na plataforma RefFinder utiliza quatro métodos de análise - geNorm (VANDESOMPELE et al., 2002), Normfinder (ANDERSON et al., 2004), BestKeeper (PFAFFL et al., 2004) e o método comparativo Delta-Ct (SILVER et al., 2006) - e, fazendo uso de média geométrica, determinam um ranking de estabilidade de genes. Na Figura 4 e na Tabela 5, são



apresentados os dados numéricos destas análises.

Figura 5: Estabilidade dos genes avaliados na plataforma RefFinder.

O algoritmo Genorm utiliza o valor M de referência, no qual genes que tiveram valores menores ou iguais a 1,5 são apropriados para uso como normalizador (VANDESOMPELE et al., 200; Silva et al., 2021). Nenhum gene dos avaliados atingiu este índice de estabilidade. Já para os demais métodos e na integração feita pelo RefFinder, ao se avaliar os dados de estabilidade considerando os três tecidos avaliados, **HSP** é o gene que apresenta os resultados mais consistentes de estabilidade. No programa NormFinder, que utiliza a variação da expressão nos grupos e entre os grupos e para estabelecer um valor de estabilidade (SV) (ANDERSON et al., 2004), o gene *HSP* foi o único a apresentar valores de

estabilidade menores que 1,0, considerado o recomendável mínimo, com valor de estabilidade 0,408.

Já na abordagem de delta-CT, que utiliza a comparação dos valores brutos de Ct como entrada para calcular a estabilidade do gene (SILVER et al., 2006), novamente o gene mais estável foi *HSP*, seguido de *UBQ* e *CYP*.

Este gene está relacionado à manutenção estrutural de proteínas em situação de estresse biótico e abiótico, no entanto as HSPs também estão presentes na ausência de estresse com função no dobramento de proteínas (UL HAQ et al., 2019). Diversos estudos utilizam *HSP* como gene normalizador em estresse abiótico (CHEN et al., 2017; SINHA et al., 2019; Wang et al., 2020), entretanto ele usualmente não se apresenta entre os mais estáveis para normalização.

Tabela 6: Estabilidade média de expressão dos genes avaliados na plataforma RefFinder

| RefFinder |         | Delta CT               |                 | BestKeeper    |       | NormFinder |   | Genorm                |      |
|-----------|---------|------------------------|-----------------|---------------|-------|------------|---|-----------------------|------|
|           |         | Média do desvio padrão |                 | Desvio padrão |       | Rankin g   |   | Valor de estabilidade |      |
| Gene      | Ranking | Média geral            | Ranking (STDEV) | Ranking       | [+/-] | g          |   | Ranking               | dade |
| HSP       | 1       | 1,09                   | 1               | 5,05          | 1     | 3,09       | 1 | 0,408                 | 1    |
| CYP       | 2       | 3                      | 3               | 6,34          | 2     | 6,07       | 3 | 5,508                 | 1    |
| UBQ       | 3       | 1,44                   | 2               | 5,87          | **    | **         | 2 | 4,567                 | 2    |

Nas análises individuais em cada tecido. CYP é o gene normalizador mais bem ranqueado para folha e raiz a partir da integralização dos dados realizados pelo RefFinder, e HSP o melhor ranqueado em caule (Tabela 3).

O gene ciclofilina (*CYP*) atua como catalisador de modificações estruturais em proteínas (SINGH et al., 2020). A Ubiquitina (*UBQ*) é classificada como um modificador polipeptídico altamente conservado em organismos eucariotos, participando ainda da degradação de proteínas. Assim como *HSP*, Ubiquitina é amplamente utilizado em análises para obtenção de genes de referência (KIARASH et al., 2018; YE et al., 2018; JIN et al., 2019).

Tabela 6: eficiência dos primers utilizados para as análises de RT-qPCR

| Gene | Eficiência do primer |
|------|----------------------|
| HSP  | 1,283                |

|     |       |
|-----|-------|
| UBQ | 1,363 |
| CYP | 1,37  |

Nas análises de eficiência, os genes apresentaram eficiência de amplificação entre 64% (*HSP*) e 69% (*CYP*) (Tabela 4). Nesse sentido, será importante a realização do teste de mais genes candidatos bem como a otimização de condições de amplificação para a adequada identificação de normalizadores de expressão.

Tabela 7: Estabilidade média de expressão dos genes em cada tecido avaliado na plataforma RefFinder

|        |      | RefFinder |             | Delta CT |                                | BestKeeper |                    | NormFinder |                       | Genorm  |                       |
|--------|------|-----------|-------------|----------|--------------------------------|------------|--------------------|------------|-----------------------|---------|-----------------------|
| Tecido | Gene | Ranking   | Média geral | Ranking  | Média do desvio padrão (STDEV) | Ranking    | Desvio padrão[+/-] | Ranking    | Valor de estabilidade | Ranking | Valor de estabilidade |
| Folha  | UBQ  | 2         | 1,86        | 1        | 0,79                           | **         | **                 | 2          | 0,707                 | 2       | 0,718                 |
|        | CYP  | 1         | 1,19        | 2        | 0,58                           | 1          | 0,44               | 1          | 0,289                 | 1       | 0,577                 |
|        | HSP  | 3         | 2,28        | 1        | 0,79                           | 2          | 0,67               | 2          | 0,707                 | 1       | 0,577                 |
| Caule  | UBQ  | 2         | 1,41        | 2        | 4,21                           | **         | **                 | 2          | 3,266                 | 1       | 2,309                 |
|        | CYP  | 3         | 3           | 3        | 5,36                           | 2          | 4,44               | 3          | 5,164                 | 2       | 4,346                 |
|        | HSP  | 1         | 1,19        | 1        | 3,46                           | 1          | 0,44               | 1          | 1,155                 | 1       | 2,309                 |
| Raiz   | UBQ  | 3         | 2,28        | 3        | 7,37                           | **         | **                 | 3          | 7,36                  | 1       | 0,577                 |
|        | CYP  | 1         | 1,19        | 1        | 3,91                           | 1          | 5,56               | 1          | 0,289                 | 1       | 0,577                 |
|        | HSP  | 2         | 1,86        | 2        | 4,04                           | 2          | 5,78               | 2          | 1,472                 | 2       | 5,106                 |

|       |                 | Ranking    |            |            |
|-------|-----------------|------------|------------|------------|
|       |                 | 1º         | 2º         | 3º         |
| Folha | Delta CT        | CYP        | UBQ        | HSP        |
|       | Bestkeeprr      | UBQ        | CYP        | HSP        |
|       | Normfinder      | CYP        | UBQ        | HSP        |
|       | Genorm          | CYP   HSP  |            | UBQ        |
|       | <b>Refinder</b> | <b>CYP</b> | <b>UBQ</b> | <b>HSP</b> |
| Caule | Delta CT        | HSP        | UBQ        | CYP        |
|       | Bestkeeprr      | UBQ        | HSP        | CYP        |
|       | Normfinder      | HSP        | UBQ        | CYP        |
|       | Genorm          | UBQ   HSP  |            | CYP        |
|       | <b>Refinder</b> | <b>HSP</b> | <b>UBQ</b> | <b>CYP</b> |
| Raiz  | Delta CT        | CYP        | HSP        | UBQ        |
|       | Bestkeeprr      | UBQ        | CYP        | HSP        |
|       | Normfinder      | CYP        | HSP        | UBQ        |
|       | Genorm          | CYP   HSP  |            | UBQ        |
|       | <b>Refinder</b> | <b>CYP</b> | <b>HSP</b> | <b>UBQ</b> |

Tabela 8: Ranqueamento dos genes de acordo com os métodos utilizados para cada tecido.

## Considerações finais

O desenvolvimento das técnicas de RT-qPCR e sequenciamento genômico trouxeram muitos avanços para o entendimento da dinâmica metabólica dos organismos vegetais. Tais dinâmicas servem como mapa para o aperfeiçoamento de técnicas relacionadas a biologia molecular e biotecnologia. Além de servir como base nos estudos descritivos sobre evolução e fisiologia de plantas.

O presente trabalho buscou acrescentar a esses estudos subsídios genômicos e transcricionais da espécie *Quillaja brasiliensis*. Para isso executamos o sequenciamento, montagem e anotação do genoma plastidial desta espécie. Como resultado, pudemos realizar uma análise filogenética com outras espécies de Fabales bem como diferenciar caracteres evolutivos referente aos cloroplastos deste grupo.

Com os dados transcricionais sequenciamos e montamos o primeiro transcriptoma de referência para a o primeiro estudo sobre seleção de genes de referência. Estes genes podem servir como guia para estudos posteriores sobre o transcriptoma desta espécie em diversas condições experimentais.

Assim o presente trabalho permitiu contribui não só para biologia da espécie *Quillaja brasiliensis* e da família Quillajaceae, mas para estudos em Fabales e em organismos produtores de Saponinas.

## Referências

ABBAS, Adeel et al. Selection and Validation of Reference Genes for RT-qPCR Analysis in *Aegilops tauschii* (Coss.) under Different Abiotic Stresses. **International journal of molecular sciences**, v. 22, n. 20, p. 11017, 2021.

ALMAS, D. Ebadi; KAMRODI, A. Rahmani. Validation of appropriate reference genes for real-time quantitative pcr gene expression analysis in rice plants exposed to metal stresses. **Russian Journal of Plant Physiology**, v. 65, n. 6, p. 890-897, 2018.

ANNADURAI, Ramasamy S. et al. De novo transcriptome assembly (NGS) of *Curcuma longa* L. rhizome reveals novel transcripts related to anticancer and antimalarial terpenoids. *PloS one*, v. 8, n. 2, p. e56217, 2013.

ANDERSEN, Claus Lindbjerg; JENSEN, Jens Ledet; ØRNTOFT, Torben Falck. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. **Cancer research**, v. 64, n. 15, p. 5245-5250, 2004.

BISWAS, Tanya; DWIVEDI, Upendra N. Plant triterpenoid saponins: biosynthesis, in vitro production, and pharmacological relevance. **Protoplasma**, v. 256, n. 6, p. 1463-1486, 2019.

BUSTIN, Stephen; NOLAN, Tania. Talking the talk, but not walking the walk: RT-qPCR as a paradigm for the lack of reproducibility in molecular research. **European Journal of Clinical Investigation**, v. 47, n. 10, p. 756-774, 2017.

BRAY, Nicolas L. et al. Near-optimal probabilistic RNA-seq quantification. **Nature biotechnology**, v. 34, n. 5, p. 525-527, 2016.

BUCCHINI, François et al. TRAPID 2.0: a web application for taxonomic and functional analysis of de novo transcriptomes. **Nucleic acids research**, v. 49, n. 17, p. e101-e101, 2021.

CIBULSKI, Samuel Paulo et al. Novel ISCOMs from *Quillaja brasiliensis* saponins induce mucosal and systemic antibody production, T-cell responses and improved antigen uptake. **Vaccine**, v. 34, n. 9, p. 1162-1171, 2016.

CHEN, Jingchao et al. Selection of relatively exact reference genes for gene expression studies in goosegrass (*Eleusine indica*) under herbicide stress. **Scientific reports**, v. 7, n. 1, p. 1-9, 2017.

CHIESA, Fernando Ferreira et al. *Quillaja brasiliensis* (A. St.-Hil. & Tul.) Mart. In: **Medicinal and Aromatic Plants of South America Vol. 2**. Springer, Cham, 2021. p. 447-459.

CHEN, Jiangfei et al. Genome-wide identification, classification and expression analysis of the HSP gene superfamily in tea plant (*Camellia sinensis*). **International Journal of Molecular Sciences**, v. 19, n. 9, p. 2633, 2018.

CHEN, Zhiao; HE, Xianghuo. Application of third-generation sequencing in cancer research. **Medical Review**, v. 1, n. 2, p. 150-171, 2021.

*Codex Alimentarius International Food Standards*; Food and Agriculture Organization of United Nations World Health Organization: Rome, Italy, 2018; Volume 49.

CUI, Yupeng et al. Full-length transcriptome analysis reveals candidate genes involved in terpenoid biosynthesis in *Artemisia argyi*. *Frontiers in genetics*, v. 12, 2021.

DA FONSECA, Rute R. et al. Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, v. 30, p. 3-13, 201

DA SILVA MAGEDANS, Yve Veronica; PHILLIPS, Michael A.; FETT-NETO, Arthur Germano. Production of plant bioactive triterpenoid saponins: From metabolites to genes and back. **Phytochemistry Reviews**, v. 20, n. 2, p. 461-482, 2021.

DAI, Yu-Lin et al. Characterization and anti-tumor activity of saponin-rich fractions of South Korean sea cucumbers (*Apostichopus japonicus*). *Journal of Food Science and Technology*, p. 1-10, 2020.

DE FARIA, Janaína Teles et al. Performance of Quillaja bark saponin and  $\beta$ -lactoglobulin mixtures on emulsion formation and stability. **Food Hydrocolloids**, v. 67, p. 178-188, 2017

DE GROOT, Carolin; MÜLLER-GOYMANN, Christel C. Saponin interactions with model membrane systems—Langmuir monolayer studies, hemolysis and formation of ISCOMs. **Planta medica**, v. 82, n. 18, p. 1496-1512, 2016.

DAI, Yu-Lin et al. Characterization and anti-tumor activity of saponin-rich fractions of South Korean sea cucumbers (*Apostichopus japonicus*). *Journal of Food Science and Technology*, p. 1-10, 2020.

DONG, Wenpan et al. ycf1, the most promising plastid DNA barcode of land plants. **Scientific reports**, v. 5, n. 1, p. 1-5, 2015.

DUAN, Z. L. et al. Reference gene selections for real time quantitative PCR analysis of gene expression in different oat tissues and under salt stress. **Biol. Plant**, v. 64, p. 838-844, 2020.

EGAN, Ashley N.; SCHLUETER, Jessica; SPOONER, David M. Applications of next-generation sequencing in plant biology. *American journal of botany*, v. 99, n. 2, p. 175-185, 2012.

FLECK, Juliane Deise et al. Saponins from *Quillaja saponaria* and *Quillaja brasiliensis*: particular chemical characteristics and biological activities. **Molecules**, v. 24, n. 1, p. 171, 2019

FENG, Shouli et al. Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in *Gossypium australe*. *BMC plant biology*, v. 19, n. 1, p. 1-16, 2019.

FIALLOS-JURADO, Jennifer et al. Saponin determination, expression analysis and functional characterization of saponin biosynthetic genes in *Chenopodium quinoa* leaves. *Plant Science*, v. 250, p. 188-197, 2016.

GEISLER, Markus; BAILLY, Aurélien; IVANCHENKO, Maria. Master and servant: Regulation of auxin transporters by FKBP and cyclophilins. ***Plant Science***, v. 245, p. 1-10, 2016.

GONZALEZ, Patricia Jurado; SÖRENSEN, Pia M. Characterization of saponin foam from *Saponaria officinalis* for food applications. *Food Hydrocolloids*, v. 101, p. 105541, 2020.

GUÉNIN, Stéphanie et al. Normalization of qRT-PCR data: the necessity of adopting a systematic, experimental conditions-specific, validation of references. *Journal of experimental botany*, v. 60, n. 2, p. 487-493, 2009.

GUZMAN, Frank et al. De novo assembly of *Eugenia uniflora* L. transcriptome and identification of genes from the terpenoid biosynthesis pathway. *Plant Science*, v. 229, p. 238-246, 2014.

HEALEY, Adam et al. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant methods*, v. 10, n. 1, p. 21, 2014.

HELLEMANS, Jan; VANDESOMPELE, Jo. Selection of reliable reference genes for RT-qPCR analysis. In: **Quantitative Real-Time PCR**. Humana Press, New York, NY, 2014. p. 19-26.

HUANG, Li-Min et al. Evolution of Terpene Synthases in Orchidaceae. *International journal of molecular sciences*, v. 22, n. 13, p. 6947, 2021.

JIN, Yuhuan et al. Identification of reliable reference genes for qRT-PCR in the ephemeral plant *Arabidopsis pumila* based on full-length transcriptome data. **Scientific reports**, v. 9, n. 1, p. 1-11, 2019.

JOSEPH, Joyous T.; POOLAKKALODY, Najya Jabeen; SHAH, Jasmine M. Plant reference genes for development and stress response studies. **Journal of biosciences**, v. 43, n. 1, p. 173-187, 2018.

JOZWIAK, Adam et al. Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery. *Nature Chemical Biology*, v. 16, n. 7, p. 740-748, 2020.

JUNG, Hyungtaek et al. Tools and strategies for long-read sequencing and de novo assembly of plant genomes. *Trends in plant science*, v. 24, n. 8, p. 700-724, 2019.

KARUNANITHI, Prema S.; ZERBE, Philipp. Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Frontiers in Plant Science*, v. 10, p. 1166, 2019.

KIARASH, Jamshidi Goharrizi et al. Selection and validation of reference genes for normalization of qRT-PCR gene expression in wheat (*Triticum durum* L.) under drought and salt stresses. **Journal of Genetics**, v. 97, n. 5, p. 1433-1444, 2018.

KIRCHHOFF, Helmut. Chloroplast ultrastructure in plants. **New Phytologist**, v. 223, n. 2, p. 565-574, 2019.

KOZERA, Bartłomiej; RAPACZ, Marcin. Reference genes in real-time PCR. **Journal of applied genetics**, v. 54, n. 4, p. 391-406, 2013.

KUMAR, Arbind et al. Plant-derived immuno-adjuvants in vaccines formulation: a promising avenue for improving vaccines efficacy against SARS-CoV-2 virus. **Pharmacological Reports**, p. 1-17, 2022.

LIAO, Pan et al. The potential of the mevalonate pathway for enhanced isoprenoid production. **Biotechnology advances**, v. 34, n. 5, p. 697-713, 2016.

LI, Rong et al. Selection of the reference genes for quantitative gene expression by RT-qPCR in the desert plant *Stipagrostis pennata*. **Scientific reports**, v. 11, n. 1, p. 1-11, 2021.

LI, Yajuan et al. Systematic identification and validation of the reference genes from 60 RNA-Seq libraries in the scallop *Mizuhopecten yessoensis*. *BMC genomics*, v. 20, n. 1, p. 1-12, 2019.

Li, Yajuan, Lingling Zhang, Ruojiao Li, Meiwei Zhang, Yangping Li, Hao Wang, Shi Wang, and Zhenmin Bao. "Systematic Identification and Validation of the Reference Genes from 60 RNA-Seq Libraries in the Scallop *Mizuhopecten Yessoensis*." *BMC Genomics* 20 (April 11, 2019): 288.

LOGSDON, Glennis A.; VOLLGER, Mitchell R.; EICHLER, Evan E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, p. 1-18, 2020.

LOVE, Michael I.; HUBER, Wolfgang; ANDERS, Simon. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, v. 15, n. 12, p. 1-21, 2014.

LUEBERT, Federico. Taxonomy and distribution of the genus *Quillaja* Molina (Quillajaceae). **Feddes Repertorium**, v. 124, n. 4, p. 157-162, 2013.

MA, Mengjie et al. Fabrication and characterization of zein/tea saponin composite nanoparticles as delivery vehicles of lutein. *LWT*, p. 109270, 2020.

MANNI, Mosè et al. BUSCO: assessing genomic data quality and beyond. **Current Protocols**, v. 1, n. 12, p. e323, 2021.

MANNI, Mosè et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. **Molecular Biology and Evolution**, v. 38, n. 10, p. 4647-4654, 2021.

MARCIANI, Dante J. Elucidating the mechanisms of action of saponin-derived adjuvants. **Trends in Pharmacological Sciences**, v. 39, n. 6, p. 573-585, 2018.

MARTICORENA, Clodomiro; MARCHESI, Eduardo. Catálogo de las plantas vasculares del Cono Sur (Argentina, Sur de Brasil, Chile, Paraguay y Uruguay).: Pteridophyta, Gymnospermae y Monocotyledoneae. **Missouri Botanical Garden Press, 2008.**

NAVARRO, E. et al. Real-time PCR detection chemistry. **Clinica chimica acta**, v. 439, p. 231-250, 2015.

NACHTIGALL, Pedro G.; KASHIWABARA, Andre Y.; DURHAM, Alan M. CodAn: predictive models for precise identification of coding regions in eukaryotic transcripts. **Briefings in bioinformatics**, v. 22, n. 3, p. bbaa045, 2021.

PRIYA, Piyush et al. Terzyme: a tool for identification and analysis of the plant terpenome. *Plant Methods*, v. 14, n. 1, p. 1-18, 2018.

PFAFFL, Michael W. et al. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. **Biotechnology letters**, v. 26, n. 6, p. 509-515, 2004.

RAMAKERS, Christian et al. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neuroscience letters*, v. 339, n. 1, p. 62-66, 2003.

RIBEIRO, Bianca et al. A Seed-Specific Regulator of Triterpene Saponin Biosynthesis in *Medicago truncatula*. *The Plant Cell*, v. 32, n. 6, 2020.

REICHERT, Corina L.; SALMINEN, Hanna; WEISS, Jochen. Quillaja saponin characteristics and functional properties. *Annual review of food science and technology*, v. 10, p. 43-73, 2019.

RUIJTER, J. M. et al. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. **Nucleic acids research**, v. 37, n. 6, p. e45-e45, 2009.

SARWAR, Muhammad Bilal et al. Identification and validation of superior housekeeping gene (s) for qRT-PCR data normalization in *Agave sisalana* (a CAM-plant) under abiotic stresses. **Physiology and Molecular Biology of Plants**, v. 26, n. 3, p. 567-584, 2020.

SEKI, Hikaru; TAMURA, Keita; MURANAKA, Toshiya. P450s and UGTs: key players in the structural diversity of triterpenoid saponins. *Plant and Cell Physiology*, v. 56, n. 8, p. 1463-1471, 2015.

SHUKLA, Pawan et al. Selection of suitable reference genes for quantitative real-time PCR gene expression analysis in Mulberry (*Morus alba* L.) under different abiotic stresses. **Molecular biology reports**, v. 46, n. 2, p. 1809-1817, 2019.

SANITÁ LIMA, Matheus et al. The (in) complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes. *Molecular Ecology Resources*, v. 16, n. 6, p. 1279-1286, 2016.

SILVA, Flávia Loyze Baldassarini et al. Validation of reference genes for real-time quantitative PCR in *Brachiaria* grass under salt stress. **Plant Gene**, v. 27, p. 100319, 2021.

SILVER, Nicholas et al. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. **BMC molecular biology**, v. 7, n. 1, p. 1-9, 2006.

SINGH, Harpreet et al. Plant cyclophilins: Multifaceted proteins with versatile roles. **Frontiers in Plant Science**, v. 11, p. 585212, 2020.

SINGH, Khushwant et al. Cyclophilins: less studied proteins with critical roles in pathogenesis. **Phytopathology**, v. 108, n. 1, p. 6-14, 2018.

SINHA, Ragini; SHARMA, T. R.; SINGH, Anil Kumar. Validation of reference genes for qRT-PCR data normalisation in lentil (*Lens culinaris*) under leaf developmental stages and abiotic stresses. **Physiology and Molecular Biology of Plants**, v. 25, n. 1, p. 123-134, 2019.

SLATKO, Barton E.; GARDNER, Andrew F.; AUSUBEL, Frederick M. Overview of next-generation sequencing technologies. **Current protocols in molecular biology**, v. 122, n. 1, p. e59, 2018.

SONI, Praveen et al. Reference gene identification for gene expression analysis in rice under different metal stress. **Journal of Biotechnology**, v. 332, p. 83-93, 2021.

STARK, Rory; GRZELAK, Marta; HADFIELD, James. RNA sequencing: the teenage years. *Nature Reviews Genetics*, v. 20, n. 11, p. 631-656, 2019.

TAYLOR, Sean et al. A practical approach to RT-qPCR—publishing data that conform to the MIQE guidelines. *Methods*, v. 50, n. 4, p. S1-S5, 2010

UL HAQ, Saeed et al. Heat shock proteins: dynamic biomolecules to counter plant biotic and abiotic stresses. **International journal of molecular sciences**, v. 20, n. 21, p. 5321, 2019.

UNTERGASSER, Andreas et al. Primer3—new capabilities and interfaces. **Nucleic acids research**, v. 40, n. 15, p. e115-e115, 2012.

VAN DIJK, Erwin L. et al. The third revolution in sequencing technology. *Trends in Genetics*, v. 34, n. 9, p. 666-681, 2018.

VANDESOMPELE, Jo et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. **Genome biology**, v. 3, n. 7, p. 1-12, 2002.

VETÖ, Nicole M. et al. Transcriptomics analysis of *Psidium cattleianum* Sabine (Myrtaceae) unveil potential genes involved in fruit pigmentation. *Genetics and molecular biology*, v. 43, 2020.

WANG, Lijuan; WANG, Yancai; ZHOU, Ping. Validation of reference genes for quantitative real-time PCR during Chinese wolfberry fruit development. **Plant physiology and biochemistry**, v. 70, p. 304-310, 2013.

WHITE, Elizabeth J. et al. Modified Cetyltrimethylammonium bromide method improves robustness and versatility: The benchmark for plant RNA extraction. Weinheim: WILEY-VCH Verlag, 2008.

WANG, Bo et al. Systematic selection and validation of suitable reference genes for quantitative real-time PCR normalization studies of gene expression in *Nitraria tangutorum*. **Scientific reports**, v. 10, n. 1, p. 1-10, 2020.

XIE, Fuliang et al. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant molecular biology*, v. 80, n. 1, p. 75-84, 2012.

YAMAMURA, Yoshimi; KUROSAKI, Fumiya; LEE, Jung-Bum. Elucidation of terpenoid metabolism in *Scoparia dulcis* by RNA-seq analysis. *Scientific reports*, v. 7,

YENDO, A. C. A. et al. Irradiance-based treatments of *Quillaja brasiliensis* leaves (A. St.-Hil. & Tul.) Mart. as means to improve immunoadjuvant saponin yield. *Industrial Crops and Products*, v. 74, p. 228-233, 2015.

YE, Jing et al. Selection of suitable reference genes for qRT-PCR normalisation under different experimental conditions in *Eucommia ulmoides* Oliv. **Scientific reports**, v. 8, n. 1, p. 1-11, 2018.

ZHANG, Zhaoping et al. Selection of the reference gene for expression normalization in *Papaver somniferum* L. under abiotic stress and hormone treatment. **Genes**, v. 11, n. 2, p. 124, 2020.

ZHAO, Yiyong et al. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. **Molecular Plant**, v. 14, n. 5, p. 748-773, 2021

ZHOU, Zhe et al. Using RNA-seq data to select reference genes for normalizing gene expression in apple roots. PloS one, v. 12, n. 9, p. e0185288, 2017.

ZHU, Jia-Hong et al. Identification, characterization and expression analysis of genes involved in steroidal saponin biosynthesis in *Dracaena cambodiana*. Journal of plant research, v. 131, n. 3, p. 555-562, 2018.

ZUO, Yang et al. The complete chloroplast genome of *Polygala japonica* Houtt.(Polygalaceae), a medicinal plant in China. Mitochondrial DNA Part B, v. 6, n. 1, p. 239-240, 2021