UNIVERSIDADE DE SÃO PAULO
FFCLRP - DEPARTAMENTO DE BIOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA COMPARADA

Dados filogenômicos para inferência de relações evolutivas entre espécies do gênero *Cereus* Mill. (Cactaceae, Cereeae)

Juliana Rodrigues Bombonato

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: BIOLOGIA COMPARADA

Ribeirão Preto - SP

2018

JULIANA RODRIGUES BOMBONATO

**Dados filogenômicos para inferência de relações evolutivas entre espécies do gênero
Cereus Mill.(Cactaceae, Cereeae)**

Orientador: Prof. Dr. Fernando de Faria Franco

Ribeirão Preto - SP

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada à fonte.

Catálogo na publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto

Bombonato, Juliana Rodrigues

Dados filogenômicos para inferência de relações evolutivas entre espécies do gênero *Cereus* Mill. (Cactaceae, Cereeae)/ Juliana Rodrigues Bombonato; orientador Fernando de Faria Franco – 2018

82 Folhas

Dissertação (Mestrado) - Programa de Pós-Graduação em Biologia Comparada, Universidade de São Paulo, Ribeirão Preto

Versão original

1. *Cereus* – 2. dados faltantes – 3. dados filogenômicos – 4. ddRADSeq – 5. NGS.

AGRADECIMENTOS

Agradeço ao meu orientador, Prof. Dr. Fernando de Faria Franco, pela oportunidade e confiança a mim depositada para realizar esse projeto, e pela ótima orientação durante todo seu desenvolvimento. O aprendizado, interesse e segurança passados foram muito importantes para meu crescimento pessoal e profissional e muitas vezes foi o que me incentivou a continuar.

Ao Prof. Dr. Evandro Marsola de Moraes, pelas contribuições e discussões durante a elaboração desse trabalho.

Agradeço ao Prof. Dr. Paulo Ribolla e ao Dr. Diego Alonso por abrirem seu laboratório para a realização de parte desse trabalho e por toda disposição em nos ajudar e aconselhar. Aproveito para agradecer ao técnico Ricardo Brassaloti, a Dra Priscilla Vilella e a Dra Karina Degaki, pelo auxílio durante as etapas do sequenciamento.

À Sonia Andrade, bioinformata que nos auxiliou em toda a preparação e triagem dos dados obtidos.

À Dra Gislaine da Silva, com quem compartilhei todas as dificuldades e conquistas durante a elaboração do protocolo de biblioteca reduzida e que foi meu apoio profissional durante essa etapa.

Agradeço a minha mãe, Renata, por todo o apoio durante essa fase e pelo amor e incentivo. Sem você não seria possível à realização desse sonho.

Agradeço aos meus irmãos Julia e Kauan, onde o bom humor e o sentimento mais puro me proporcionaram um refúgio nos momentos de turbulência. Agradeço a minha avó, Maria, por todo amor e carinho, além do aconchego que recebo toda vez que por ela procuro.

Aos meu primos e primas, por sempre estarem ao meu lado e pelos ótimos momentos que passamos juntos.

Agradeço ao meu companheiro, Wesley, pelo amor, pela companhia e apoio em todos os momentos, principalmente os de crises. Você foi fundamental para o sucesso deste projeto.

Aos colegas e ex-colegas de laboratório e da UFSCar, Heidi Utsunomia, Juliana Martinez, Isabel Bonatelli, Danilo Amaral, Manolo Perez, Gislaine da Silva, Monique Romeiro, Fernanda Mizuguchi, Gulzar Khan, Felipe Silva, Fernanda Mayumi, Milena Telhe e Danielle

Bittencourt, com quem aprendi pessoal e profissionalmente. Obrigado pela convivência, amizade, ensinamento e pelos momentos de descontração. Vocês foram parte essencial neste caminho.

À USP *campus* Ribeirão Preto, ao Programa de Pós-graduação em Biologia Comparada (FFCLRP) e a secretária desse programa, Vera de Lucca, por todo apoio prestado.

À UFSCar *campus* Sorocaba e ao Laboratório de Diversidade Genética e Evolução (LaGEvol – Departamento de Biologia), por permitirem que eu desenvolvesse esse trabalho em suas instalações.

Às agências de fomento à pesquisa FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo – Projeto nº 2014/25227-0) e CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro aqui investido.

“The cacti represent the most spectacular New World radiation of succulent plants”
(ARAKAKI et al. 2011)

RESUMO

BOMBONATO, Juliana Rodrigues. **Dados filogenômicos para inferência de relações evolutivas entre espécies do gênero *Cereus* Mill. (Cactaceae, Cereeae).** 82 FOLHAS. Dissertação (Mestrado em Ciências) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (Departamento de Biologia), Universidade de São Paulo, Ribeirão Preto, 2018.

Estudos filogenômicos usando Sequenciamento de Próxima Geração (do inglês, *Next Generation Sequencing* - NGS) estão se tornando cada vez mais comuns. O uso de marcadores oriundos do sequenciamento de DNA de uma biblioteca genômica reduzida, neste caso ddRADSeq (do inglês, *Double Digestion Restriction Site Associated DNA Sequencing*), para este fim é promissor, pelo menos considerando sua relação custo-benefício em grandes conjuntos de dados de grupos não-modelo, bem como a representação genômica recuperada. Aqui usamos ddRADSeq para inferir a filogenia em nível de espécie do gênero *Cereus* (Cactaceae). Esse gênero compreende em cerca de 25 espécies reconhecidas predominantemente sul-americanas distribuídas em quatro subgêneros. Nossa amostra inclui representantes de *Cereus*, além de espécies dos gêneros próximos, *Cipocereus* e *Praecereus*, além de grupos externos. A biblioteca ddRADSeq foi preparada utilizando as enzimas *EcoRI* e *HPAII*. Após o controle de qualidade (tamanho e quantificação dos fragmentos), a biblioteca foi sequenciada no Illumina *HiSeq* 2500. O processamento de bioinformática a partir de arquivos FASTQ incluiu o controle da presença de adaptadores, filtragem por qualidade (softwares FastQC, MultiQC e SeqyClean) e chamada de SNPs (software iPyRAD). Três cenários de permissividade a dados faltantes foram realizados no iPyRAD, recuperando conjuntos de dados com 333 (até 40% de dados perdidos), 1440 (até 60% de dados perdidos) e 6141 (até 80% de dados faltantes) loci. Para cada conjunto de dados, árvores de Máxima Verossimilhança (MV) foram geradas usando duas supermatrizes: SNPs ligados e Loci. Em geral, observamos algumas inconsistências entre as árvores ML geradas em softwares distintos (IQTree e RaxML) ou baseadas no tipo de matriz distinta (SNPs ligados e Loci). Por outro lado, a precisão e a resolução, foram melhoradas usando o maior conjunto de dados (até 80% de dados perdidos). Em geral, apresentamos uma filogenia com resolução inédita para o gênero *Cereus*, que foi resolvido como um provável grupo monofilético, composto por quatro clados principais e com alto suporte em suas relações internas. Além disso, nossos dados contribuem para agregar informações sobre o debate sobre o aumento de dados faltantes para conduzir a análise filogenética com loci RAD.

Palavras-chave: *Cereus*, dados faltantes, dados filogenômicos, ddRADSeq, NGS.

ABSTRACT

BOMBONATO, Juliana Rodrigues. **Dados filogenômicos para inferência de relações evolutivas entre espécies do gênero *Cereus* Mill. (Cactaceae, Cereeae).** 82 FOLHAS. Dissertação (Mestrado em Ciências) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (Departamento de Biologia), Universidade de São Paulo, Ribeirão Preto, 2018.

Phylogenomics studies using Next Generation Sequencing (NGS) are becoming increasingly common. The use of Double Digest Restriction Site Associated DNA Sequencing (ddRADSeq) markers to this end is promising, at least considering its cost-effectiveness in large datasets of non-model groups as well as the genome-wide representation recovered in the data. Here we used ddRADSeq to infer the species level phylogeny of genus *Cereus* (Cactaceae). This genus comprises about 25 species recognized predominantly South American species distributed into four subgenera. Our sample includes representatives of *Cereus*, in addition to species from the closely allied genera *Cipocereus* and *Praecereus*, besides outgroups. The ddRADSeq library was prepared using *EcoRI* and *HPAII* enzymes. After the quality control (fragments size and quantification) the library was sequenced in Illumina *HiSeq* 2500. The bioinformatic processing on raw FASTQ files included adapter trimming, quality filtering (FastQC, MultiQC and SeqyClean softwares) and SNPs calling (iPyRAD software). Three scenarios of permissiveness to missing data were carry out in iPyRAD, recovering datasets with 333 (up tp 40% missing data), 1440 (up to 60% missing data) and 6141 (up to 80% missing data) loci. For each dataset, Maximum Likelihood (ML) trees were generated using two supermatrices: SNPs linked and Loci. In general, we observe few inconsistencies between ML trees generated in distinct softwares (IQTree and RaxML) or based in distinctive matrix type (SNP linked and Loci). On the other hand, the accuracy and resolution were improved using the larger dataset (up to 80% missing data). Overall, we present a phylogeny with unprecedented resolution for genus *Cereus*, which was resolved as a likely monophyletic group, composed by four main clades and with high support in their internal relationships. Further, our data contributes to aggregate information on the debate about to increasing missing data to conduct phylogenetic analysis with RAD loci.

Keywords: *Cereus*, missing data, phylogenomic data, ddRADSeq, NGS.

LISTA DE FIGURAS

- Figura 1** - Imagem destacando as relações dentro do clado BCT redesenhada com base na filogenia de Cactaceae proposta por Hernández-Hernández et al. (2014).....15
- Figura 2** - Filogenia colapsada estabelecida por Inferência Bayesiana, baseada em segmentos do espaçador intergênico plastidial *trnS-trnG*. As probabilidades posteriores (> 0.85), os valores de suporte de ramos para Máxima Verossimilhança (> 60) e os valores de Parcimônia (> 60) são sequencialmente indicados perto dos ramos do nó. Fonte: FRANCO et al. (2017a).....17
- Figura 3** - Análise quantitativa dos fragmentos entre 400pb e 700pb, realizada em um *Bioanalyzer* (Agilent Genomics), onde os picos de 11 a 15 representam o intervalo de tamanho mais representativo, sendo o tamanho médio de 576pb. Os picos de 35pb e 10380pb são referentes aos marcadores.....38
- Figura 4** - Matriz de ocupação no cenário de até 40% de dados faltantes indicando presença e ausência dos loci em cada amostra (Tabela 1). Em vermelho estão destacadas amostras do *outgroup*. Os quadrados em preto indicam presença enquanto os brancos indicam ausência.....43
- Figura 5** - Matriz de ocupação no cenário de até 60% de dados faltantes indicando presença e ausência dos loci em cada amostra (Tabela 1). Em vermelho estão destacadas amostras do *outgroup*. Os quadrados em preto indicam presença enquanto os brancos indicam ausência.....44
- Figura 6** - Matriz de ocupação no cenário de até 80% de dados faltantes indicando presença e ausência dos loci em cada amostra (Tabela 1). Em vermelho estão destacadas amostras do *outgroup*. Os quadrados em preto indicam presença enquanto os brancos indicam ausência.....45
- Figura 7** - Esquema mostrando os suportes de ramos para os principais clados das filogenias geradas para *Cereus* neste trabalho (Figura 8; Figura 9, APÊNDICES de D a K), de acordo com os diferentes programas, medidas de suporte, matrizes de dados e cenários de dados faltantes. Onde, a cor verde indica suporte significativo e a cor amarela indica suporte não-significativo para cada uma das medidas de suporte de ramos (SH-aLRT > 80 , *ultrafast bootstrap* > 95 e *rapid bootstrap* > 70).....46

Figura 8 - Árvore filogenética colapsada estabelecida por Máxima Verossimilhança baseada na matriz SNPs ligados contendo 96.097 sítios e permitindo até 80% de dados faltantes gerada nos programas IQTree e RaXML. Os suportes de ramos foram considerados significativos quando maiores que 80 para Shimodaira-Hasegawa + aLRT (SH-aLRT), maiores que 95 para o *ultrafast bootstrap* (UFBOOT) e maiores que 70 para o *rapid bootstrap* (RBS), representados nessa ordem na árvore acima dos ramos. O * indica suporte de 100 para ambas as medidas (SH-aLRT, UFBOOT, RBS) e os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura. Os quadrados coloridos indicam as áreas de ocorrência das espécies do gênero *Cereus*.....47

Figura 9 - Árvore filogenética colapsada estabelecida por Máxima Verossimilhança baseada na matriz Loci contendo 6141 loci e permitindo até 80% de dados faltantes gerada nos programas IQTree e RaXML. Os suportes de ramos foram considerados significativos quando maiores que 80 para Shimodaira-Hasegawa + aLRT (SH-aLRT), maiores que 95 para o *ultrafast bootstrap* (UFBOOT) e maiores que 70 para o *rapid bootstrap* (RBS), representados nessa ordem na árvore acima dos ramos. O * indica suporte de 100 para ambas as medidas (SH-aLRT, UFBOOT, RBS) e os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura. Os quadrados coloridos indicam as áreas de ocorrência das espécies do gênero *Cereus*.....48

LISTA DE TABELAS

Tabela 1 - Descrição das amostras utilizadas na execução do presente trabalho, identificando qual amostra estava presente em cada um dos protocolos de biblioteca reduzida desenvolvidos.....	26
Tabela 2 - Números de leituras geradas pelos sequenciamentos (<i>MiSeq</i> e <i>HiSeq</i>) e após as filtrações (SeqClean e iPyRAD), além do número de loci gerados em cada um dos cenários de dados faltantes antes (iPyRAD) e após a retirada das duplicatas (CD-HIT).....	39
Tabela 3 - Número de bases em cada matriz de dados, número de sítios informativos e constantes e porcentagem de CG para cada um dos cenários de dados faltantes.....	42

SUMÁRIO

1	INTRODUÇÃO	14
1.1	A família Cactaceae	14
1.2	Modelo Biológico: o gênero <i>Cereus</i> (Cactaceae, Cereae)	16
1.3	Obtenção de marcadores moleculares para inferências filogenéticas	18
2	HIPÓTESES DO PRESENTE TRABALHO	24
3	OBJETIVO	25
4	METODOLOGIA	26
4.1	Amostragem	26
4.2	Simulações de digestões <i>IN SÍLICO</i>	29
4.3	Protocolo ddRADSeq (Biblioteca II)	30
4.3.1	Digestão de DNA genômico e Ligação de Adaptadores	30
4.3.2	Seleção de tamanho, PCR e Pool	31
4.3.3	Controle de qualidade da biblioteca e Sequenciamento	32
4.3.4	Análises de Bioinformática	33
4.4	Análises Filogenéticas	34
4.4.1	Matrizes de dados	34
4.4.2	Abordagens de Máxima Verossimilhança (MV)	35
5	RESULTADOS	38
5.1	Protocolo ddRADSeq (Biblioteca II)	38
5.2	Análises filogenéticas	42
6	DISCUSSÃO	50
6.1	Considerações a respeito de diferentes protocolos ddRADSeq	50
6.2	Análises filogenéticas	52

7 CONCLUSÕES.....	59
REFERÊNCIAS.....	60
APÊNDICES.....	69
APÊNDICE A.....	69
APÊNDICE B.....	73
APÊNDICE C.....	74
APÊNDICE D.....	75
APÊNDICE E.....	76
APÊNDICE F.....	77
APÊNDICE G.....	78
APÊNDICE H.....	79
APÊNDICE I.....	80
APÊNDICE J.....	81
APÊNDICE K.....	82

1. INTRODUÇÃO

1.1 A família Cactaceae

A família Cactaceae, amplamente distribuída em todo continente americano, é extremamente diversa, com aproximadamente 100 gêneros e mais de 1850 espécies descritas (HUNT; TAYLOR; CHARLES, 2006; BARTHLOTT et al. 2015), distribuídas em três subfamílias, Pereskioideae, Opuntoideae e Cactoideae, das quais a última é a mais diversa com cerca de 80% das espécies da família (NYFFELER, 2002). A diversidade dos cactos em relação a suas formas é fascinante, podendo ir de árvores a arbustos, ramificados ou não ramificados, apresentando cladódios suculentos colunares ou globulares (NYFFELER, 2002).

São conhecidos sete centros com maior riqueza de espécies desta família no continente Americano (BARTHLOTT et al. 2015), dos quais o leste do Brasil (região Nordeste e maioria do Sudeste) é um dos principais, o que é razoável, uma vez que essa região abriga habitats característicos para o estabelecimento dessas plantas, como a Caatinga, o Cerrado e seus respectivos ecótonos com a Mata Atlântica (ZAPPI; TAYLOR; LAROCCA, 2011). De toda a grande variedade de espécies de Cactaceae, cerca de 130 espécies são endêmicas do leste do Brasil, e esse número pode se tornar ainda maior em termos de endemismo se considerarmos o Brasil Central e o Brasil Meridional (ZAPPI; TAYLOR; SANTOS, 2011). A família Cactaceae é considerada o 5º grupo taxonômico mais ameaçado mundialmente, onde cerca de 31% de suas espécies nominais estão classificadas com algum grau de ameaça na Lista Vermelha da União Internacional para a Conservação da Natureza e dos Recursos Naturais (*IUCN Red List*) (GOETTSCHE et al. 2015).

Trabalhos de sistemática molecular envolvendo os táxons de Cactaceae, apontam que sua origem ocorreu na transição Eoceno-Oligoceno (aproximadamente 35 milhões de anos atrás) na região andina do Chile, da Argentina e da Bolívia, porém toda sua diversidade surgiu nos últimos 10 milhões de anos na transição Mioceno-Plioceno, o que é congruente com a expansão dos biomas áridos nesta época no continente americano (ARAKAKI et al. 2011; HERNÁNDEZ-HERNÁNDEZ et al. 2011).

Na filogenia da família proposta por Hernández-Hernández et al. (2011), um subclado da subfamília Cactoideae foi denominado BCT (composto pelas tribos Cereeae e Trichocereae)

(Figura 1). O clado BCT apresenta uma das maiores taxas de diversificação da família, tendo sua distribuição predominantemente associada ao continente sul-americano e início de sua diversificação datada entre 6 e 5 milhões de anos atrás. A origem e diversificação desse clado, assim como de toda a família Cactaceae, provavelmente está associada não só com o aumento da aridificação, mas também com baixas temperaturas, diminuição da precipitação, queda global dos níveis de CO₂ registrados desde o Eoceno e evolução de sistemas polinizadores (ARAKAKI et al. 2011; HERNÁNDEZ-HERNÁNDEZ et al. 2011).

Figura 1 - Imagem destacando as relações dentro do clado BCT redesenhada com base na filogenia de Cactaceae proposta por Hernández-Hernández et al. (2014).

A ocupação dos ambientes áridos pelos cactos, foi favorecida por várias especializações encontradas nos representantes dessa família, conhecidas como síndrome de suculência, sendo elas: (1) potencial mais eficiente de armazenamento água baseado em seu mecanismo fotossintético especializado CAM (*Crassulacean Acid Metabolism*) que permite absorção de gás carbônico durante a noite e minimizando a perda de água; (2) modificação das folhas em espinhos; (3) caule com pigmentos fotossintéticos e cutícula espessa na epiderme; (4) produção de tricomas nas raízes e, (5) tecidos que permitem a estocagem de líquidos (NYFFELER, 2002;

ARAKAKI et al. 2011). As hipóteses e as características apresentadas acima, tornam os táxons da família Cactaceae interessantes modelos biológicos tanto para estudos de diversificação rápida e recente quanto para estudos que investiguem mudanças climáticas do Quaternários sobre a biota Neotropical (MORAES et al. 2005; MAJURE et al. 2012; BONATELLI et al. 2014; MENEZES et al. 2016, PEREZ et al. 2016a).

1.2 Modelo Biológico: o gênero *Cereus* (Cactaceae, Cereae)

O gênero *Cereus* Mill. (Cactaceae; Cereae) é um dos gêneros de Cactaceae mais representados no leste do Brasil, compreendendo 25 espécies reconhecidas com ocorrência associada majoritariamente à América do Sul (HUNT; TAYLOR; CHARLES, 2006), mas com algumas espécies ocorrendo em países da América Central, como *Cereus haitiensis* (FRANCK et al. 2017). Filogeneticamente, o gênero *Cereus* está alocado no clado BCT de Cactaceae (Figura 1).

Esse gênero está subdividido em quatro subgêneros (*Cereus*, *Mirabella*, *Ebneria* e *Oblongicarpus*) (TAYLOR; ZAPPI, 2004; HUNT; TAYLOR; CHARLES, 2006). O subgênero *Cereus* é o mais diverso com 13 espécies (*C. stenogonus*, *C. bicolor*, *C. hildmannianus*, *C. jamacaru*, *C. hexagonus*, *C. pierre-braunianus*, *C. trigonodendron*, *C. fernambucensis*, *C. insularis*, *C. hankeanus*, *C. cochabambensis*, *C. vargasianus*, *C. lanosus*, *C. lamprospermus*) ocorrendo na maioria dos biomas sul-americanos, tendo no Brasil sua distribuição associada a ambientes de Caatinga, Cerrado, Mata Atlântica, Pampa e Pantanal. Os subgêneros *Mirabella* (*C. mirabella* e *C. aubicaulis*) e *Ebneria* (*C. aethiops*, *C. spegazzinni*, *C. kroenleinii*, *C. saddianus*, *C. estevesii*, *C. phatnospermus*) ocorrem em vegetações do Cerrado no leste brasileiro e no sudoeste da América do Sul, respectivamente. Finalmente, o subgênero *Oblongicarpus* (*C. fricii*, *C. repandus*, *C. mortensenii*, *C. horrispinus*) que ocorre no extremo norte da América do Sul (HUNT; TAYLOR; CHARLES, 2006), na Colômbia, Venezuela e Antilhas. Importante salientar, que muitas variedades do gênero *Cereus* estão descritas como subespécies, como por exemplo *C. jamacaru* subsp. *jamacaru* e *C. jamacaru* subsp. *calcirupicula*, o que aumenta o número de táxons descritos.

No trabalho de Franco et al. (2017a), dados moleculares do espaçador intergênico cloroplastidial *trnS-trnG* foram usados para inferir as relações de parentesco entre diversas espécies de *Cereus* e indicou que esse gênero pode não ser monofilético, uma vez que os

gêneros *Cipocereus* e *Praecereus* foram posicionados internamente entre as espécies de *Cereus*. Além disso, todos os subgêneros de *Cereus* mostraram-se polifiléticos neste trabalho, com exceção do subgênero *Mirabella* (Figura 2). A origem de *Cereus* foi datada durante a transição Plioceno-Pleistoceno (4,24 - 0,95 milhões de anos atrás), tendo o Cerrado como provável área ancestral, e a maioria dos eventos de diversificação ocorrendo durante o Pleistoceno médio (FRANCO et al. 2017a).

Figura 2 - Filogenia colapsada estabelecida por Inferência Bayesiana, baseada em segmentos do espaçador intergênico plastidial *trnS-trnG*. As probabilidades posteriores (> 0.85), os valores de suporte de ramos para Máxima Verossimilhança (> 60) e os valores de Parcimônia (> 60) são sequencialmente indicados perto dos ramos do nó. Fonte: FRANCO et al. (2017a).

Além das questões de posicionamentos de táxons, é possível também observar nessa filogenia a baixa resolução e baixos valores de suporte para os nós internos (Figura 2). De fato, a falta de resolução em filogenias de Cactaceae com base em marcadores de cloroplasto é um resultado comumente encontrado (BUTTERWORTH; WALLACE, 2004; RITZ et al. 2007; CALVENTE et al. 2011; FRANCK; COCHRANE; GAREY, 2013; CALVENTE et al. 2017), revelando a dificuldade desse marcador para resolver relações de parentesco nessa família. Esse tipo de resultado na família é frequentemente discutido considerando o modelo de diversificação recente e explosivo do grupo (RITZ et al. 2007; VÁZQUEZ-SÁNCHEZ et al.

2013), o que contribui para a manutenção do polimorfismo ancestral nas linhagens evolutivas emergentes e tendo como outra consequência filogenética comprimento de ramos longos e relações não estabelecidas (FRANCO et al. 2017a).

Um outro estudo envolvendo espécies de *Cereus* também baseado em poucos marcadores moleculares (gene nuclear fitocromo C (PHYC) e no espaçador intergênico *trnS-trnG*) (FRANCO et al. 2017b), usou as espécies do subclado A2 (*C. fernambucensis* e *C. insularis*) (Figura 2) para discutir os eventos que moldaram a vegetação em ambientes xéricos da Mata Atlântica e investigar a estrutura populacional e os mecanismos que levaram diversificação dessas espécies. Quatro grupos genéticos foram encontrados dentro dessas duas espécies, e uma íntima relação filogenética entre as populações do norte da distribuição de *C. fernambucensis* com *C. insularis*, tornando *C. fernambucensis* um grupo parafilético (Franco et al. 2017b).

As hipóteses anteriores levantadas para as espécies de *Cereus* (Franco et al. 2017a e Franco et al. 2017b) precisariam ser testadas com um conjunto maior de marcadores, uma vez que árvores baseadas em apenas um ou poucos loci, foram muito utilizadas nos últimos 30 anos e são eficazes em muitas abordagens, porém em muitos casos são eficientes apenas para uma primeira avaliação, ficando esse tipo de análise aquém do desejado devido a incongruência entre árvores de espécies e árvores de genes, causada por processos como manutenção de polimorfismo ancestral, transferência horizontal e duplicação/extinção gênica (MADDISON, 1997; LELIAERT et al. 2014), assim o aumento no número de marcadores moleculares tem sido sugerido como estratégia para a obtenção de topologias mais semelhantes entre árvores de genes e árvores de espécie e também para a obtenção de maiores valores de suporte de ramos (EDWARDS, 2008; LELIAERT et al. 2014; DACOSTA; SONRENSON, 2015; HOSNER et al. 2015).

1.3 Obtenção de Marcadores Moleculares para inferências filogenéticas

Os marcadores moleculares baseiam-se na existência de variabilidade e na possibilidade de sua detecção, sendo que para inferência filogenética marcadores moleculares genealógicos precisam ser capazes de traçar a história das relações evolutivas entre alelos ou haplótipos. Para tanto algumas características são desejáveis, como presença de muitas cópias no genoma, revelar certo grau de polimorfismo, mas ao mesmo tempo ter uma taxa de mutação que o

permita ser compartilhado entre espécies focais, além de ser de fácil e rápido isolamento (SMALL; CRONN; WENDEL, 2004; PETIT; EXCOFFIER, 2009; SHAW et al. 2014). Para o estabelecimento de filogenias moleculares no nível de espécie marcadores de DNA plastidial (mitocôndria e cloroplasto) e regiões não codificantes do DNA nuclear foram amplamente exploradas desde a década de 80 (AVISE, 2009; AVISE; BOWEN; AYALA, 2016, ALLENDORF, 2017).

Porém, a busca por marcadores com variabilidade genética suficiente para abordar questões em níveis taxonômicos mais superficiais, ou seja, dentro e entre espécies, permitindo a inferência de eventos históricos ou recorrentes que moldaram essas relações evolutivas, continuou sendo um grande desafio na biologia molecular (SCHAAL; OLSEN, 2000; AVISE, 2009; TURCHETTO-ZOLET et al. 2013; SHAW et al. 2014), que se torna ainda mais desafiador em um contexto de espécies de divergência recente. A resolução de filogenias em grupos de divergência rápida e recente são mais laboriosas, uma vez que o pouco tempo de divergência entre os táxons dificultam a identificação de linhagens, o que pode estar relacionado com processos de coalescência profunda/manutenção de polimorfismo ancestral entre táxons, assim marcadores tradicionalmente usados podem não ter informação suficiente para esclarecer relações de parentesco nesses grupos (MADDISON, 1997; BELFIORE; LIU; MORITZ, 2008).

Em plantas, marcadores moleculares oriundos do genoma cloroplastidial (cpDNA) são comumente utilizados para estabelecimento de hipóteses filogenéticas entre espécies de divergência recente (SHAW et al. 2005), uma vez que evolui de 3 a 4 vezes mais rápido do que o mtDNA em plantas, além de sua elevada quantidade de cópias em plantas, herança geralmente uniparental e a disponibilidade de diversos “primers universais” (AVISE, 2009; SHAW et al. 2014). No entanto, também existem limitações no uso de cpDNA como marcador filogenético, como sua maior suscetibilidade a introgressão (SMALL; CRONN; WENDEL, 2004; PETIT; EXCOFFIER, 2009), e a grande variação nas taxas evolutivas das regiões do cpDNA mesmo entre espécies de divergência recente (SHAW et al. 2014; ROMEIRO-BRITO et al. 2016). Além disso, estudos recentes com cactos revelaram que marcadores oriundos do genoma nuclear apresentam maior variabilidade em relação aos plastidiais (ROMEIRO-BRITO et al. 2016; PEREZ et al. 2016a).

Marcadores nucleares são considerados os mais informativos para investigações em baixos níveis taxonômicos (SHAW et al. 2014), uma vez que a sua organização em diferentes cromossomos permite que esse genoma englobe loci submetidos a diferentes condições e, por

isso, apresente genealogias distintas que juntas carregam assinaturas sobre a história evolutiva de um grupo, porém seu uso é restringido devido ao ritmo relativamente lento de evolução das sequências (como exceção tem-se os íntrons em genes codificadores de proteínas que evoluem de forma mais rápida em relação ao mtDNA e ao cpDNA), a presença de recombinação intergênica e a dificuldade de se isolar haplótipos de organismos diploides (AVISE, 2009).

Além das considerações sobre taxas e modos de evolução, o entendimento da Teoria da Coalescência mostrou que as árvores de genes são realizações aleatórias de um processo estocástico, de modo que a história de um gene pode não refletir a real história de uma espécie, tendo profundas implicações na análise e interpretação de dados (MADDISON, 1997). Dessa forma, percebeu-se que uma maior resolução filogenética poderia ser alcançada com o uso de abordagens multilocos envolvendo amostragens de diferentes genomas, incorporando nas análises a evolução independente de diferentes regiões genômicas, o que aumenta a chance de resgatar a história evolutiva de um táxon (BELFIORE; LIU; MORITZ, 2008; EDWARDS, 2008; BRITO; EDWARDS, 2009; NIELSEN; BEAUMONT, 2009).

Até pouco tempo a obtenção de muitos marcadores moleculares, principalmente de regiões nucleares, para estudos filogenéticos era uma tarefa dispendiosa e financeiramente custosa, o que dificultava o estabelecimento de filogenias multilocos (ELSHIRE et al. 2011; HARVEY et al. 2013; DIJK et al. 2014). Nos últimos anos, o chamado Sequenciamento de Nova Geração (do inglês, *Next Generation Sequencing*- NGS), capaz de sequenciar milhões de pares de bases de DNA em pouco tempo, tem permitindo a obtenção de regiões genômicas mesmo em organismos não-modelos e sem um genoma de referência (STRAUB et al. 2012; ZIMMER et al. 2012; DOYLE, 2013; SOUSA et al. 2014).

O uso dessa tecnologia foi relativamente atrasado durante algum tempo, apesar de rentável e vantajoso, para estudos com finalidades filogenéticas e filogeográficas devido ao uso, nesses trabalhos, de organismos não-modelos, grande número de amostras por espécies e falta de consenso entre os protocolos de preparação de bibliotecas reduzidas. Atualmente esse cenário está sendo modificado e o uso do NGS tem permitindo a obtenção de marcadores moleculares em diversos organismos e para diversos objetivos (MCCOMARCK et al. 2013; DACOSTA; SONRENSON, 2015). Isso foi possível graças a metodologias capazes de reduzir genomas, acessando apenas uma parte de toda a informação genética dos organismos por meio de sondas e/ou enzimas de restrição, assim possibilitando o desenvolvimento de centenas de marcadores moleculares distribuídos pelos genomas mitocondrial, cloroplastidial e nuclear (i.e.

EMERSON et al. 2010; WAGNER et al. 2013; BONATELLI; CARSTENS; MORAES, 2015; PEREZ et al. 2016b; DUPUIS et al. 2017).

As tecnologias NGS apesar de extremamente rentáveis em relação a seu custo/benefício, apresentam dificuldades que merecem ser destacadas. Um dos primeiros pontos a se considerar é a questão da cobertura durante o sequenciamento, ou seja, quantas vezes cada fragmento será sequenciado, o que irá depender do tamanho do genoma, do número de amostras e do tipo de protocolo de redução utilizado. A baixa cobertura pode trazer problemas durante processamento de dados, sendo que diferentes estratégias para redução de genoma podem minimizar esse problema (MCCORMACK et al. 2013; ANDREWS et al. 2016). Outro fato que pode comprometer tal tipo de estudo está ligado a estocasticidade durante o sequenciamento, onde não existe um controle exato do que está sendo sequenciado. Uma outra característica de algumas tecnologias NGS de 2ª geração, é sua limitação em relação ao comprimento dos fragmentos gerados. A maioria das plataformas Illumina geram fragmentos de até 300pb em sequenciamentos *paired-end*, considerados fragmentos curtos (MCCORMACK et al. 2013), o que pode representar um problema para alguns trabalhos, porém as plataformas de 3ª e 4ª geração (PacBio e Nanopore, respectivamente) já produzem fragmentos maiores (RHOADS; AU, 2015; FENG et al. 2015)

Algumas estratégias NGS têm sido usadas em estudos filogenéticos, dentre elas: métodos de captura de sequências (SeqCap) adjacentes a regiões genômicas ultra-conservadas (UCEs) e a preparação de bibliotecas reduzidas por digestão do DNA genômico com enzimas de restrição (*Restriction Site Associated DNA Sequencing* - RADSeq; *Genotyping by Sequencing* - GBS) (DAVEY; BLAXTER 2010; ELSHIRE et al. 2011; LA HARPE et al. 2017). Apesar da primeira abordagem ter algumas vantagens para estudos filogenéticos envolvendo táxons de divergência antiga em relação ao segundo (HARVEY et al. 2013), essa metodologia tende a recuperar loci associados a regiões sob forte influência de seleção purificadora, podendo estar sujeitos ao efeito carona, e comprometendo o sinal filogenético. Os métodos de redução do genoma, por outro lado, permitem maior amplitude na exploração da amostragem genômica, o que supostamente é uma vantagem para recuperar a aleatoriedade do processo coalescente nos diferentes loci associados a diferentes regiões (EDWARDS, 2008; LIU et al. 2009), além de serem o mais utilizado em estudos filogenéticos e filogeográficos (MCCOMARCK et al. 2013; ANDREWS et al. 2016).

Em um recente trabalho, La Harpe et al. (2017) destacou o potencial de métodos RadSeq, no estudo de radiação de espécies estreitamente relacionadas, onde essas metodologias representariam fontes valiosas para obtenção de milhares de marcadores genéticos para estudar radiações. De fato, o uso dessa metodologia para realização de filogenias tem sido ampliado na literatura (MCCOMARCK et al. 2013; DACOSTA; SONRENSON, 2015; DUPUIS et al. 2017), mesmo para grupos que iniciaram a divergência a cerca de 60 milhões de anos atrás (RUBIN; REE; MOREAU, 2012).

A biblioteca ddRADSeq (do inglês *Double digestion restriction site-associated DNA sequencing*) foi descrita pela primeira vez por Peterson (2012) e é a abordagem genômica mais amplamente utilizada em filogenias (ANDREWS et al. 2016), que ao invés de uma (como no RADSeq) utiliza duas enzimas de restrição na etapa de digestão do DNA durante o preparo da biblioteca reduzida. Esse método é capaz de genotipar milhares de SNPs (do inglês, *Single Nucleotide Polymorphism*) sequenciando regiões curtas adjacentes aos locais de corte das enzimas de restrição e não é necessário um genoma de referência, o que possibilita a aplicação em espécies não-modelo (ANDREWS et al. 2014; PURITZ et al. 2014).

O uso de duas enzimas de restrição é uma boa alternativa para genomas grandes, já que contribui para uma maior redução de seu tamanho (PETERSON et al. 2012; PURITZ et al. 2014), além da diminuição do custo e a possibilidade do uso de muitos indivíduos (PURITZ et al. 2014). Porém, a alta quantidade de DNA alvo em boa qualidade pode representar um desafio no uso desse método (ANDREWS et al. 2014), além da etapa de seleção de tamanho manual, que pode levar a representação variável de alguns loci, levando a menos fragmentos ortólogos entre os indivíduos. Para este último, equipamentos para seleção automática (por exemplo, Pippin Prep) já são capazes de realizar a seleção de tamanho de forma mais precisa (MCCOMARCK et al. 2013; PURITZ et al. 2014). As duplicatas de PCR, que são leituras clonais do mesmo fragmento de DNA genômico geradas durante a etapa de PCR na preparação da biblioteca reduzida, podem “transformar” um heterozigoto em um homozigoto ou fazer com que um alelo contendo um erro de PCR pareça um alelo real (ANDREWS et al. 2014) e para evitar ou diminuir este efeito encontrado em bibliotecas ddRADSeq, os autores indicam a preparação e sequenciamento de fragmentos *paired-end*, múltiplas PCRs e também o uso de bons softwares e bons critérios durante o a chamada de SNPs (ANDREWS et al. 2014; PURITZ et al. 2014; ANDREWS et al. 2016).

Além dos já citados, a melhor maneira de se trabalhar com o grande conjunto de dados gerados por essas tecnologias é o novo desafio que está gerando conflitos na literatura. Essa questão envolve alguns pontos como, por exemplo, a exclusão ou inclusão de dados faltantes, uma vez que em bibliotecas reduzidas, dados podem ser perdidos em muitas etapas, o que resultará na presença ou ausência de um determinado loco no conjunto final de loci, afetando o tamanho da matriz de dados, o tempo computacional durante as análises e os suportes de ramos na árvore filogenética (HUANG; KNOWLES, 2014; HOSNER et al. 2015). Como e quais dados gerados devem ser utilizados (ex.: apenas SNPs ou sua combinação com sítios invariantes, indels, presença-ausência de sítios de restrição), é outro ponto que vem sendo investigado e considera possíveis ruídos que cada conjunto de dados pode trazer para as análises filogenéticas, como o aumento do comprimento de ramos no caso de matrizes contendo apenas sítios variáveis, que pode ser intensificado pelo aumento de dados faltantes (DACOSTA; SONRENSEN et al. 2015; LEACHÉ et al. 2015).

2. HIPÓTESES DO PRESENTE TRABALHO

Com base no conhecimento atual sobre a filogenia de *Cereus* podem ser estabelecidas algumas hipóteses: 1) o gênero não seria monofilético, diante da inclusão de espécies dos gêneros *Cipocereus* e *Praecereus* como ramificações internas e seria composto por quatro linhagens evolutivas principais (A-D; Figura 2); 2) o subgênero *Mirabella* seria o único monofilético dentre os subgêneros de *Cereus* (Figura 2); 3) mesmo sem suporte na filogenia baseada no DNA plastidial, é possível estabelecer uma hipótese de que as espécies dos clados A e B de *Cereus* (Figura 2), que incluem as espécies arbóreas e de maior porte do gênero popularmente denominadas de “mandacaru”, compõem um grupo monofilético; 4) a espécie *C. fernambucensis* seria parafilética devido a íntima relação filogenética entre as populações do norte da distribuição de *C. fernambucensis* com *C. insularis* (Franco et al. 2017b) e; 5) considerando a ideia de diversificação rápida e recente das espécies de *Cereus* (Franco et al. 2017a), é esperado que com o aumento da variação genética filogeneticamente informativa auxilie na compreensão tanto de relações mais profundas como de relações terminais do gênero.

3. OBJETIVO

Esse trabalho é parte de um projeto maior que busca o entendimento da história evolutiva das espécies do gênero *Cereus*. Especificamente, o principal objetivo deste trabalho é desenvolver uma hipótese filogenética robusta para o gênero *Cereus*, com resolução tanto nas relações profundas como nos terminais do grupo, permitindo um melhor entendimento dos processos de diversificação e levantamento de futuras hipóteses populacionais e biogeográficas para o grupo. Para tanto alguns objetivos específicos foram traçados, sendo eles:

- (1) padronizar um protocolo de preparação de biblioteca reduzida (ddRADSeq) para espécies do gênero *Cereus*;
- (2) obter um conjunto de marcadores moleculares informativos para o gênero *Cereus*;
- (3) verificar o impacto da presença de dados faltantes e do tipo de dado utilizado (sítios variantes e invariantes), nas análises filogenéticas; e
- (4) estabelecer uma hipótese filogenética para o gênero *Cereus* com base nos loci obtidos no presente projeto e discutir os resultados com base nas hipóteses previamente estabelecidas com dados de DNA plastidial (Figura 2).

4. METODOLOGIA

4.1 Amostragem

As amostras selecionadas para a preparação da biblioteca genômica reduzida estão listadas na Tabela 1. As amostras de *Cereus* foram escolhidas com base nos táxons usados para estabelecimento da hipótese filogenética preliminar de *Cereus* (Figura 2). Uma amostragem maior foi definida para a espécie *C. fernambucensis*, considerando que essa espécie é composta por pelo menos quatro grupos populacionais e provavelmente parafilética, devido ao posicionamento de *C. insularis* como grupo irmão das populações mais ao norte da distribuição de *C. fernambucensis* (FRANCO et al. 2017b). Além disso, foram incluídas espécies dos gêneros *Cipocereus* e *Praecereus* diante do seu posicionamento interno na filogenia de *Cereus* (Figura 2). Para a amostragem de *Praecereus euchlorus* foi realizada uma coleta na região de Votorantim, SP (23°36'41.2''S; 47°23'59.1''W) (Tabela 1). Finalmente, espécies de diversos gêneros do clado BCT foram incluídos na amostragem como grupos externos. Dentre esses, com exceção das amostras de *Micranthocereus auriazureus*, *Pilosocereus arrabidaei* e *Uebelannia pectinífera*, todos os demais foram obtidas por importação de DNA genômico de plantas em cultivo no *Kew Royal Botanic Gardens* (Licença CITES nº 16BR020235/DF), conforme descrito na Tabela 1.

Para a extração de DNA, amostras de raiz de cada um dos indivíduos da Tabela 1, foram maceradas em nitrogênio líquido, com o objetivo de romper as paredes celulares, seguido da extração com os reagentes do *DNeasy Plant Mini Kit* (Qiagen). Para algumas amostras foram realizadas mais de uma extração para que fosse possível atingir a massa necessária de DNA (290 ng) para realização das etapas de preparação das bibliotecas reduzidas. As extrações foram visualizadas em gel de agarose 2% e as soluções de DNA quantificadas por fluorimetria no *Qubit 3.0 Fluorometer* (Invitrogen) com o *QubitTM dsDNA HS Assay Kit*.

Tabela 1- Descrição das amostras utilizadas na execução do presente trabalho, identificando qual amostra estava presente em cada um dos protocolos de biblioteca reduzida desenvolvidos.

ESPÉCIE	LOCALIDADE GEOGRÁFICA	CÓDIGO	VOUCHER	BIBLIOTECA
INGROUP				
subgênero <i>Cereus</i>				
<i>C. bicolor</i>	Rondonópolis, MT	S103E1	SORO 4570	I/II
	Jauru-Coxim, MS	S52V3	SORO 2743	I/II
	Dianópolis, TO	S144F2	T	I/II

Tabela 1. Continua

ESPÉCIE	LOCALIDADE GEOGRÁFICA	CÓDIGO	VOUCHER	BIBLIOTECA
<i>INGROUP</i>				
subgênero <i>Cereus</i>				
<i>C. cochabambensis</i>	Cochambamba - Bolívia	S152A10	32072	II
	Ubatuba, SP	S72F11	SORO 2658	I/II
		S72F12	SORO 2658	I/II
	Arraial do Cabo, RJ	S80F8	SORO 2663	I/II
		S80F9	SORO 2663	II
<i>C. fernambucensis</i>		S114F8	SORO 2675	I
<i>fernambucensis</i>	Una, BA	S114F9	SORO 2675	II
		S114F15	SORO 2675	II
		S114F16	SORO 2675	I
	Maracajaú, RN	S104F4	SORO 2670	I/II
		S104F17	SORO 2671	I/II
	Três Rios, RJ	S76F6	SORO 2662	I/II
		S76F10	SORO 2662	I/II
<i>C. fernambucensis</i>	Bom Jesus do Norte, RJ	S85F3	SORO 2745	I/II
<i>sericifer</i>		S85F2	SORO 2745	I/II
	Águia Branca, MG	S88F3	SORO 2749	I/II
		S88F4	SORO 2749	I/II
	São João do Manteninhas, MG	S77A24	SORO 2775	I/II
<i>Cereus hankeanus</i>	SEM LOCALIDADE	S152A6	32068	II
<i>Cereus hexagonus</i>	SEM LOCALIDADE	S152A9	32071	II
	Salto de Itu, SP	S90V3	SORO 2746	I/II
<i>C. hildmannianus</i>	Poconé, MT	S103C1	SORO 4566	I
		S103C	SORO 4566	II
	Fernando de Noronha, PE (Praia do Boldró)	S115C41	SORO 2677	I/II
<i>C. insularis</i>	Fernando de Noronha, PE (Trilha Forte do Sueste)	S115B21	SORO 2678	I/II
	Fernando de Noronha, PE (Praia do Atalaia)	S115A1	SORO 2677	I
		S115A8	SORO 2677	II
<i>C. jamacaru calcirupicola</i>	Capitão Enéas, MG	S44V3	SORO 4487	I/II
	São Desiderio, BA	S38V7B	SORO 2740	I/II
		S113F5	SORO 5702	II
<i>C. jamacaru jamacaru</i>	Conceição de Feira, BA	S113F7	SORO 5702	I/II
		S113F9	SORO 5702	I

Tabela 1. Continuação

ESPÉCIE	LOCALIDADE GEOGRÁFICA	CÓDIGO	VOUCHER	BIBLIOTECA
INGROUP				
subgênero <i>Cereus</i>				
<i>C. jamacaru jamacaru</i>	Penedo, AL	S110F4	SORO 4567	I/II
<i>C. pierre-braunianus</i>	Aurora do Tocantins, TO	S77A7	SORO 2741	I/II
		S145F4	SORO 4565	I/II
<i>C. stenogonus</i>	Bolívia	S149V11	T	I/II
	Porto Murtinho, MS	S149V15	T	I/II
	SEM LOCALIDADE	S152A12	32074	II
<i>Cereus trigonodendron</i>	Tarapoto - Peru	S152A4	32066	II
subgênero <i>Ebneria</i>				
<i>C. aethiops</i>	NA	S77A30	SORO 2774	I/II
<i>C. kroenleinii</i>	Cipolândia, MS	S131F4	SORO 4563	I/II
<i>C. saddianus</i>	Cáceres, MT	S103D	SORO 3632	II
		S103D1	SORO 3632	I
<i>C. spgazzinni</i>	Rondonópolis, MT	S99A6	SORO 4564	I/II
	Porto Murtinho, MS	S149V14	T	I/II
	NA	S77A31	SORO 4541	I/II
subgênero <i>Mirabella</i>				
<i>C. albicaulis</i>	Uburama, BA	S77A17	SORO 2771	I/II
	Morro do Chapéu, BA	S38V6B	SORO 4559	I/II
<i>C. mirabella</i>	Barreiras, BA	S77A18	SORO 2778	I
		S151V3	SORO 2779	II
	Chapada Gaúcha, MG	S147F6	SORO 4560	I/II
subgênero <i>Oblongicarpi</i>				
<i>Cereus fricii</i>	SEM LOCALIDADE	S152A8	32070	II
<i>Cipocereus</i>				
<i>Cipocereus bradei</i>	Augusto de Lima, MG	S137A1	SORO 4562	I/II
<i>Cipocereus minensis</i>	Presidente Kubitschek, MG	S10B6	SORO 4560	I
		S151A1	SORO 4561	II
<i>Praecereus</i>				
<i>Praecereus euchlorus</i>	Votorantim, SP	S150F1	SORO 5704	I/II
		S150F2	SORO 5704	I/II

Tabela 1. Conclusão

ESPÉCIE	LOCALIDADE GEOGRÁFICA	CÓDIGO	VOUCHER	BIBLIOTECA
INGROUP				
<i>Praecereus</i>				
<i>Praecereus saxicola</i>	Porto Murtinho, MS	S149V16	SORO 5703	I/II
OUTGROUP				
<i>Browningia hertlingiana</i>	SEM LOCALIDADE	S152A7	32069	II
<i>Echinopsis formosa</i>	SEM LOCALIDADE	S152A15	32077	II
<i>Espositoopsis dybowski</i>	Nosso Sr do Bomfim, BA	S152A13	32075	II
<i>Gymnocalycium denudatum</i>	SEM LOCALIDADE	S152A3	32065	II
<i>Haageocereus pseudomelanostele</i>	SEM LOCALIDADE	S152A2	32064	II
<i>Matucana intertexta</i>	Matara, Cajanarca - Peru	S152A1	32063	II
<i>Micranthocereus auriazureus</i>	Grão Mogol I, MG	S26A16	HUFS647	I/II
<i>Oreocereus celsianus</i>	Culpina - Bolívia	A152A5	32067	II
<i>Pilosocereus arrabidaei</i>	Nova Viçosa, BA	S35M1A	SORO 4488	I/II
<i>Samaipaticereus corroanus</i>	SEM LOCALIDADE	S152A14	32076	II
<i>Uebelmannia pectinifera</i>	Mendanha 3	S62A19	SORO 4543	I
Biblioteca I- protocolo segundo Kess et al. 2016 Biblioteca II- protocolo segundo Campos et al. 2017 T - em tombo				

4.2 Simulações de digestões *IN SÍLICO*

Diferentes pares de enzimas de restrição podem ser aplicados para a construção da biblioteca genômica (YANG et al. 2016). Neste trabalho realizamos simulações *in silico* em um pacote do R denominado SimRAD (LEPAIS; WEIR et al. 2014), o qual nos possibilitou a escolha do par de enzimas que foi aplicado no protocolo de biblioteca reduzida.

A escolha dos pares de enzimas testados se baseou na sua aplicação em Cactaceae, como é o caso do par *EcoRI* + *MseI* (Perez et al. 2016b) e em trabalhos mais amplos com diferentes grupos de angiospermas (YANG et al. 2016). Em ambos os casos se considerou o número de fragmentos gerados e a cobertura que cada fragmento teria no sequenciamento, de acordo com o intervalo de tamanho de seleção dos fragmentos. Para predição de um genoma simulado, foi

fornecido como informação prévia ao programa o tamanho médio do genoma de Cactaceae (2.5 Gb; NÉGRON-ORTIZ, 2007) e a porcentagem de CG (40%) encontrada no genoma de *Cereus*, que foi obtida com o resultado de um sequenciamento do genoma parcial de um indivíduo de *C. fernambucensis*, realizado paralelamente a este projeto (Franco FF e colaboradores, dados não publicados). A digestão *in silico* do genoma simulado foi realizada para os pares de enzimas *EcoRI* + *MseI*, *EcoRI* + *HpaII*, *PstI* + *MspI*, *AvaII* + *MspI*. Uma observação importante deve ser feita, uma vez que Yang et al. (2016) utilizou a enzima *MspI*, sendo esta equivalente a enzima *HPAII* utilizada neste trabalho, diferindo apenas na sensibilidade para a metilação de citosina (XIA et al. 2014).

Para seleção de tamanho também foram observadas as simulações realizadas no SimRAD buscando avaliar qual seria o número de loci esperado simulando diferentes intervalos de seleção de tamanho e a respectiva cobertura no sequenciamento. Os valores gerados foram usados para calcular a cobertura esperada no sequenciamento de uma biblioteca com 63 indivíduos.

4.3 Protocolo ddRADSeq (Biblioteca II)

Primeiramente, é necessário destacar que foram preparados duas bibliotecas reduzidas neste trabalho, a Biblioteca I (segundo protocolo de Kess et al. (2016)) e a Biblioteca II (segundo Campos et al. (2017)) (Tabela 1). Porém, o protocolo da Biblioteca I apresentou resultados insatisfatórios que impossibilitaram o uso desses dados para as futuras análises, com isso será apresentado neste tópico apenas o desenvolvimento detalhado do protocolo da Biblioteca II, enquanto os dados da Biblioteca I estão registrados no Apêndice A.

4.3.1 Digestão de DNA genômico e Ligação de Adaptadores

Foi preparada uma biblioteca reduzida com 63 indivíduos da Tabela 1. Para tanto foi utilizado o protocolo ddRADSeq descrito em Campos et al. (2017), que consiste em uma modificação do protocolo desenvolvido por Peterson et al. (2012). Esse protocolo foi realizado no Laboratório de Pesquisa em Análises Genéticas no Instituto de Biotecnologia (IBTEC) da UNESP *campus* Botucatu em colaboração com o Prof^o Dr Paulo Eduardo M. Ribolla.

Para este protocolo utilizamos alíquotas da solução de DNA genômico de indivíduos da Tabela 1, sendo que um total de 290ng do DNA genômico desses indivíduos foi então digerido simultaneamente com 5U de cada uma das enzimas [*EcoRI* (20U/μL) e *HPAII* (10U/μL)]. As digestões ocorreram por 3h a 37°C e posteriormente, as amostras foram purificadas com *beads* magnéticas (Agencourt® AMPure® XP), sendo recuperado um volume de 35μl para cada amostra. O DNA purificado foi quantificado no *Qubit 3.0 Fluorometer* (Invitrogen) com o *Qubit™ dsDNA HS Assay Kit*.

Para a reação de ligação, os adaptadores fita-dupla (APÊNDICE B) foram ligados ao produto de digestão por meio de uma reação com 0,4μM do adaptador de cada uma das enzimas, 31,7μL do produto de digestão, 4μL de T4 *Ligase* Buffer 1X, 0,6U da enzima T4 DNA *Ligase* e água ultrapura para completar 40μL de volume. Essa reação foi incubada por 30min a 23°C, seguida pela inativação de 10 minutos a 65°C.

4.3.2 Seleção de tamanho, PCR e Pool

As simulações no SimRAD para o par de enzimas *EcoRI* + *HPAII* indicaram como melhor intervalo de seleção sendo entre 300pb e 400pb, gerando uma cobertura potencial de aproximadamente 11x, ou seja, 11 leituras por fragmento. Portanto, foram excisados do gel fragmentos entre 400pb a 500bp, considerando o acréscimo de aproximadamente 120bp nos produtos de digestão após a ligação dos adaptadores. A seleção de tamanho foi realizada em gel de agarose 1%, corado com o intercalante *SYBR™ Safe DNA Gel Stain* (Invitrogen). O marcador de peso molecular utilizado foi o 1 kb *DNA Ladder* (NEB, Inc) e *Low Mass DNA Ladder* (Invitrogen). O fragmento selecionado do gel para cada amostra, foi congelado por 1h no freezer -20°C, centrifugado por 10 min a 14000 rpm, purificado com *beads* magnéticas (Agencourt® AMPure® XP) e eluído em um volume final de 35uL.

Para inserir a sequência dos identificadores (indexes) i5 e i7 padronizados pela Illumina (APÊNDICE B), bem como uma sequência complementar a “*flow-cell*” Illumina, nos insertos ligados aos adaptadores, foi utilizada uma PCR de fusão com *primers* contendo sequencias complementares aos adaptadores *EcoRI* e *HPAII*, além dos indexes i5 e i7 e a sequência da *flow-cell* Illumina. Para tanto, foi montada uma reação contendo 4μL do produto de ligação, 1μL de cada primer de fusão e 10μL de *Phusion High-Fidelity PCR Master Mix* (aproximadamente 0,3 U de *Phusion*) (Thermo Scientific). Os primers foram combinados em

cada reação de forma que cada amostra tivesse uma combinação única. As reações de PCR foram realizadas em triplicata visando aumentar a concentração de DNA final e minimizar a amplificação preferencial de alelos (PETERSON et al. 2012). A ciclagem da reação de PCR envolveu 17 ciclos com um passo inicial de 72°C por 3 minutos, dois passos de desnaturação (95°C por 30 segundos e 95°C por 10), um passo de anelamento (55°C por 30 segundos), um passo de extensão (72°C por 30 segundos) e um passo de extensão final a 72°C por 5 minutos.

Ao final, as triplicatas foram combinadas e purificadas com *beads* magnéticas (Agencourt® AMPure® XP), recuperando um volume final de 30µL por amostra. As amostras já purificadas foram quantificadas no Qubit 2.0 *dsDNA HS Assay Kit*, e a partir dessa quantificação as amostras foram normalizadas para aproximadamente 4 ng/µL e uma solução de *pool* com todas as amostras em condições equimolares foi preparada.

4.3.3 Controle de qualidade da biblioteca e Sequenciamento

A etapa de controle de qualidade tem por objetivo garantir e confirmar a qualidade da biblioteca em relação ao seu tamanho e concentração para entrada no sequenciamento e, para tanto, o *pool* gerado após a PCR foi analisado por meio do *Bioanalyzer* (Agilent Genomics) e de PCR em tempo real (qPCR). A análise no *Bioanalyzer* permitiu visualizar, de forma quantitativa o espectro de tamanho selecionado, bem como visualizar picos indicativos de excesso de adaptadores, um problema comum em bibliotecas RadSeq. A qPCR possibilitou a análise da concentração apenas dos fragmentos ligados a adaptadores da “*flow-cell*” Illumina, indicando se a massa da biblioteca seria suficiente para entrar no sequenciamento. Antes do sequenciamento final, foi realizado um sequenciamento preliminar da biblioteca na plataforma *MiSeq* Illumina com *MiSeq Reagent Nano Kit v2* (2 x 150 pb – aproximadamente 1 milhão de *reads*) para observação do resultado em uma menor escala. O sequenciamento final da biblioteca ddRadSeq foi realizado no Centro de Genômica Funcional ESALQ-USP em um sequenciador *HiSeq 2500* no modo de corrida rápida, utilizando *HiSeq RapidRun Flow Cell v2*, com o *HiSeq Rapid SBS Kit v2* e leituras pareadas de 125 bp (2x).

4.3.4 Análises de Bioinformática

Para as análises de bioinformática foi inserido à amostragem do *outgroup* um indivíduo de *Uebelmannia pectiífera* (Mendanha-3 - Diamantina-MG, S62A21, SORO 4543) preparado da mesma forma descrita e sequenciado em um trabalho paralelo do nosso grupo de pesquisa.

As análises de controle de qualidade dos dados das foram realizadas nos programas FastQC (Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>) e MultiQC (Disponível em: <https://github.com/ewels/MultiQC>) antes e após os passos de filtragem do programa SeqyClean 1.9.12 (ZHBANNIKOV et al. 2017). Ambos os programas de controle de qualidade, realizam análises estatísticas para vários indicadores de controle de qualidade, porém o primeiro faz essa análise para cada amostra, enquanto o segundo combina os resultados do FastQC e apresenta os resultados para o conjunto de amostras. Para a remoção das sequencias de adaptadores e outros contaminantes que poderiam estar presentes na biblioteca, foi utilizado o software SeqyClean 1.9.12, com a seguinte linha de comando:

```
./seqyclean -v UniVec.fas -qual 20 20 -minlen 65 -1 INPUT_R1.fastq -2 INPUT_R2.fastq -o NOME DO OUTPUT
```

Como pode ser observado acima, foi utilizado um arquivo que contém, além de sequências de adaptadores Illumina, sequências de muitos outros contaminantes que podem ser encontrados em bibliotecas de DNA, o UniVec.fas, que foi utilizado como uma base de dados para o programa SeqyClean e pode ser encontrado no banco de dados UniVec (Disponível em: <https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/>). A qualidade mínima (-qual) exigida para a permanência da sequência foi QScore = 20 (*Phred Score*), valor frequentemente usado para eliminar sequências de baixa qualidade (MCCOMARCK et al. 2013) e o comprimento mínimo (-minlen) em pares de bases exigido para manter uma sequência no conjunto final de dados foi 65. Após a “limpeza” dos dados os arquivos resultantes dos sequenciamentos *MiSeq* e *HiSeq* foram combinados para que os dados do sequenciamento *MiSeq* também pudessem ser aproveitados além apenas de um controle de qualidade.

O pacote iPyRAD (Disponível em: <http://github.com/dereneaton/ipyrad>) foi usado para a chamada de SNPs após a filtragem no SeqyClean, de acordo com alguns parâmetros como similaridade entre as sequências dentro e entre amostras, cobertura mínima e controle de sequencias parálogas. Essa análise ocorreu em três cenários distintos, de acordo com a cobertura mínima de amostras em que o loco precisaria estar presente para ser mantido no conjunto final

de loci. Essa abordagem tem por objetivo avaliar a influência dados faltantes em reconstruções filogenéticas. Assim, o mesmo conjunto de dados foi triado por três vezes no iPyRAD, com os seguintes cenários de dados faltantes: até 40% (loco presente em no mínimo 39 amostras), até 60% (loco presente em no mínimo 26 amostras), e até 80% (loco presente em no mínimo 13 amostras), de dados faltantes para a manutenção de um loco na filtragem. Para todos os cenários foi utilizado uma filtragem estrita, a qual filtra as sequências por qualidade (QScore = 20), presença de sequências semelhantes a adaptadores Illumina e tamanho mínimo da sequência. Mesmo considerando a alta qualidade das sequências (QScore = 30) após a filtragem no SeqyClean observada nos resultados do MultiQC, um filtro estrito foi escolhido no iPyRAD por apresentar melhores resultados de agrupamento (“clusterização”) das sequências (dentro e entre indivíduos) em relação a filtros mais permissivos. Além disso, também foi exigido um nível de 85% de similaridade entre as sequências, dentro e entre as amostras, considerando a distância genética entre os táxons amostrados. O arquivo de parâmetros dessas análises pode ser observado no APÊNDICE C.

Após a triagem dos loci, foi utilizado o programa CD-HIT (LI; GODZIK, 2006; FU et al. 2012) que buscou por sequências iguais consideradas diferentes (chamadas aqui de sequências duplicadas) durante a chamada de SNPs no iPyRAD, o que pode superestimar o número de loci isolados e, conseqüentemente, impactar nas análises filogenéticas. Quando encontradas, as sequências duplicadas foram retiradas sempre optando por manter aquelas com maior comprimento em pares de base.

4.4 Análises Filogenéticas

4.4.1 Matrizes de dados

Para cada conjunto de dados (até 40%, 60% e 80% de dados faltantes) foram preparadas duas matrizes de dados concatenados (super matriz) para realização das análises filogenéticas, sendo elas: Loci (SNPs + sítios invariáveis) e SNPs ligados (sequências apenas com sítios variáveis). Aqui é importante ressaltar que a denominação “SNPs ligados” não tem relação com ligação gênica e, apenas se refere ao fato de todos os sítios variantes de cada loci estarem ligados sem serem separados por sítios invariantes, sendo essa uma denominação muito comum nos programas de processamento de SNPs.

A matriz de Loci foi obtida através da manipulação do arquivo .gphocs (*output* iPyRAD), após a retirada de sequências duplicadas e posterior concatenação dos arquivos contendo as sequências de cada loci no programa AMAS (BOROWIEC, 2016), gerando arquivos concatenados nos formatos NEXUS e Phylip. A matriz de SNPs ligados foi obtida a partir da manipulação do arquivo NEXUS, com sequências contendo sítios variáveis + sítios invariáveis (matriz Loci), para retirada de todos os invariantes, em um pacote no R denominado Phrynomics 2.0 (Disponível em: <https://github.com/bbanbury/phrynomics>), sendo este um conjunto de ferramentas para manipulação de dados de SNPs.

4.4.2 Análise de Máxima Verossimilhança (MV)

Primeiramente, a amostra S152A12 da espécie *C. stenogonus* (Tabela 1) foi retirada das análises, uma vez que em análises filogenéticas preliminares essa amostra foi sempre alocada juntamente com os grupos externos, colocando dúvida sua correta identificação como pertencente ao gênero *Cereus*. Importante ressaltar que a origem dessa amostra foi o *Kew Gardens*, de onde a maior parte dos grupos externos foi obtida.

A análise de Máxima Verossimilhança concatenada foi implementada no RaXML 8 (STAMATAKIS, 2014) e no IQTree 1.5.6 (NGUYEN et al. 2015). Embora seja um ponto de discussão, a decisão pela concatenação nos dados desse tipo tem se justificado pela maioria dos estudos empíricos na literatura assumindo a premissa de que o volume de sinal filogenético no grande conjunto de dados se sobrepõe às potenciais discordâncias entre as partições (WAGNER et al. 2013; DACOSTA; SONRENSON et al. 2015; DUPUIS et al. 2017).

Os programas escolhidos realizam diferentes algoritmos para escolha da melhor árvore, sendo desejável o uso de ambos para avaliação de possíveis impactos do algoritmo de reconstrução nas árvores geradas (REDDY et al. 2017; ZHOU et al. 2017). Além disso, o RaXML e o IQTree apresentam medidas de *bootstrap* distintas, permitindo uma avaliação do suporte de ramos mais acurada. Os suportes de *bootstrap*, comumente empregados em análises filogenéticas, é uma técnica que infere valores de confiança em árvores filogenéticas a partir da reconstrução de muitas árvores com pequenas variações entre elas e devido a sua alta exigência computacional e frente aos novos e grandes conjuntos de dados genéticos disponíveis, seu uso

tornou-se dispendioso, o que vem exigindo renovações deste modelo (STAMATAKIS; HOOVER; ROUGEMONT, 2008).

O *bootstrap* rápido (*rapid bootstrap*, RBS) (STAMATAKIS; HOOVER; ROUGEMONT, 2008) disponível em versões mais recentes do programa RaXML é mais rápido que o *bootstrap* tradicional (*standart bootstrap*, SBS), uma vez que explora menos o espaço amostral, sendo indicado para conjunto de dados de até aproximadamente 8000 sequências, atuando até 20 vezes mais rápido e obtendo resultados similares aos fornecidos pelo SBS (STAMATAKIS; HOOVER; ROUGEMONT, 2008). O *ultrafast bootstrap* (UFBoot) (HOANG et al. 2017), adotado no IQTree, além de explorar menos o espaço amostral, combinou o método de log-verossimilhança com uma coleta mais eficaz de árvores candidatas e também possui um algoritmo que avalia a convergência de valores de suporte de ramificação para determinar automaticamente quando parar de colecionar árvores candidatas, tornando o UFBoot até 30 vezes mais rápido que o RBS (MINH et al. 2013; HOANG et al. 2017). Além das diferenças de tempo de execução e exigência computacional, o suporte de ramo é outro ponto distinto entre esses *bootstraps*, onde o UFBoot utiliza um suporte de ramo considerável a partir de 95%, enquanto no RBS valores acima de 70% são muitas vezes adotados (i.e. WAGNER et al. 2013; REDDY et al. 2017).

Além dos suportes de *bootstrap*, também foi utilizado, apenas no IQTree, um segundo suporte de ramos conhecido como SH-aLRT (GUINDON et al. 2010), que é uma versão não-paramétrica do teste aproximado de razão de verossimilhança (aLRT) usando a comparação de múltiplas árvores realizada pelo teste Shimodaira-Hasegawa (SH) (ANISIMOVA; GASCUEL, 2006), sendo um teste de suporte de ramo simples mais rápido do que todas as análises de *bootstrap* e recomendado para conjuntos de dados extremamente grandes. O suporte de ramo deste algoritmo deve ser considerado significativo a partir de 80 (GUINDON et al. 2010).

Além disso, para a matriz compostas apenas de sítios variáveis, em ambos os programas, foi utilizado uma correção conhecida como modelo ASC (LEWIS, 2001). Essa correção, originalmente descrita para estimar filogenias morfológicas com dados discretos, vem sendo usada para filogenias moleculares quando matrizes de dados contendo apenas SNPs (sem os sítios invariantes) são utilizadas, corrigindo a probabilidade condicionada em sítios variáveis e sem ela o comprimento dos ramos podem ser superestimados afetando a seleção da topologia das árvores (LEWIS, 2001; LEACHÉ et al. 2015).

Em todas as análises de suporte realizadas foram utilizadas 1000 réplicas, seja de *bootstrap* ou SH-aLRT. A escolha do modelo de substituição nucleotídica utilizado nas análises filogenéticas foi feita através da análise pelo software JModelTest 2.1.10 (POSADA, 2008) utilizando a matriz de Loci sem os grupos externos, sendo que para ambas as implementações, o modelo GTR + G (GTRGAMMA) foi recuperado, entre os 11 modelos testados, como o modelo de evolução que melhor explicaria nossos dados segundo os critérios de Akaike (AIC) e Akaike corrigido para pequenas amostras (AICc).

Todas as análises do IQTree foram realizadas em um CPU Dell core i7 (8 processadores e 24 Gb de memória RAM), assim como a maioria das análises do RaXML, exceto as matrizes de SNPs ligados e Loci para o cenário de até 80% de dados faltantes, nestes casos, devido ao grande volume de dados, as análises foram realizadas no Cluster Maritaca – UFSCar Sorocaba (16 *nodes* e 32 Gb de memória RAM) e no servidor SGI (196 nodes e 1.5 Tb de memória RAM), respectivamente.

5. RESULTADOS

5.1 Protocolo ddRADSeq (Biblioteca II)

A solução de *pool* com todas as amostras em condições equimolares foi analisada em um *Bioanalyzer* que indicou o tamanho entre 400pb e 700pb, onde o intervalo entre 480 e 620pb contém a maior parte em concentração nesta biblioteca e com média de 576pb. Descontando os adaptadores flaqueadores (~120pb) seriam esperados insertos entre 350pb e 490pb sendo sequenciados. Tanto a análise quantitativa dos fragmentos (Figura 3) quanto a concentração estimada na qPCR (13,80 nM), foram resultados considerados satisfatórios para realização do sequenciamento massivo.

Figura 3 - Análise quantitativa dos fragmentos entre 400pb e 700pb, realizada em um *Bioanalyzer* (Agilent Genomics), onde os picos de 11 a 15 representam o intervalo de tamanho mais representativo, sendo o tamanho médio de 576pb. Os picos de 35pb e 10380pb são referentes aos marcadores.

A análise de qualidade realizada no FastQC e do MultiQC não mostraram contaminação por adaptadores Illumina, sendo que a qualidade das sequências se encontrava, em sua maioria, acima de QScore = 30 exceto para o início das seqüências, o que já era esperado devido a característica intrínseca de bibliotecas ddRADSeq, de baixa complexidade no início das seqüências devido ao compartilhamento das bases do sítio de restrição das enzimas.

O sequenciamento *MiSeq* nano gerou 1.581.822 leituras e o *HiSeq RapidRun* resultou em 308.567.198 leituras. Os dados resultantes desses sequenciamentos foram analisados em conjunto de modo *paired-end* (PE) e podem ser observados na Tabela 2.

Tabela 2- Números de leituras geradas pelos sequenciamentos (*MiSeq* e *HiSeq*) e após as filtrações (SeqClean e iPyRAD), além do número de loci gerados em cada um dos cenários de dados faltantes antes (iPyRAD) e após a retirada das duplicatas (CD-HIT).

Amostras	Nº leituras analisadas (não pareadas)			Nº leituras após o SeqClean (pareadas)	Nº leituras após o iPyRAD (pareadas)	Nº Loci (iPyRAD)			Nº Loci (CD-HIT)		
	<i>MiSeq</i>	<i>HiSeq</i>	<i>MiSeq + HiSeq</i>	<i>MiSeq + HiSeq</i>	<i>MiSeq + HiSeq</i>	40%	60%	80%	40%	60%	80%
S103E1	24.222	2.972.578	2.996.800	1.202.023	1.202.017	259	1243	3134	245	1110	2733
S52V3	27.022	3.144.434	3.171.456	1.298.723	1.298.682	314	1120	2574	297	1005	2255
S144F2	24.360	2.800.536	2.824.896	1.159.836	1.159.826	287	1178	2738	270	1060	2457
S152A10	36.680	3.841.900	3.878.580	1.625.841	1.625.814	275	1070	3032	258	953	2640
S72F11	19.522	2.449.520	2.469.042	1.002.480	1.002.478	229	1173	3336	213	1043	2866
S72F12	15.552	1.903.914	1.919.466	774.539	774.538	288	1039	2968	269	930	2548
S80F9	9.908	1.118.474	1.128.382	460.571	460.571	207	726	1634	197	659	1452
S80F8	20.170	2.493.734	2.513.904	1.013.543	1.013.543	237	1017	2645	225	928	2379
S114F15	22.728	2.499.070	2.521.798	1.036.050	1.036.041	284	1195	3113	266	1062	2705
S114F9	25.204	2.875.722	2.900.926	1.199.838	1.199.823	332	1302	3747	312	1155	3226
S104F4	32.910	3.305.982	3.338.892	1.396.236	1.396.170	280	1229	3881	263	1093	3311
S104F17	33.956	3.406.180	3.440.136	1.442.600	1.442.588	317	1117	3177	296	1003	2781
S76F6	38.034	3.954.138	3.992.172	1.662.772	1.662.767	265	1238	3877	249	1100	3301
S76F10	77.762	10.181.672	10.259.434	4.135.525	4.135.504	313	1372	4699	293	1225	4058
S85F3	40.190	4.465.416	4.505.606	1.857.355	1.857.350	339	1304	3932	318	1179	3459
S85F2	17.768	2.634.464	2.652.232	1.030.880	1.030.873	260	1136	3433	247	1013	2944
S88F3	1.436	217.040	218.476	80.873	80.873	99	183	238	99	172	224
S88F4	14.298	2.063.806	2.078.104	806.854	806.846	281	934	2327	263	848	2069
S77A24	12.450	1.694.376	1.706.826	681.995	681.995	315	1109	2900	295	988	2514
S152A6	38.060	3.923.634	3.961.694	1.662.283	1.662.272	294	981	2701	278	886	2372
S152A9	36.224	3.693.650	3.729.874	1.547.762	1.547.747	320	1151	3197	301	1018	2701
S90V3	22.380	3.041.596	3.063.976	1.233.112	1.233.045	309	1109	3187	290	989	2720
S103C	16.768	2.124.584	2.141.352	864.099	864.069	258	1130	2731	244	1013	2350

Tabela 2. Continua

Amostras	Nº leituras analisadas (não pareadas)			Nº leituras após o SeqyClean (pareadas)	Nº leituras após o iPyRAD (pareadas)	Nº Loci (iPyRAD)			Nº Loci (CD-HIT)		
	<i>MiSEQ</i>	<i>HiSeq</i>	<i>MiSeq + HiSeq</i>	<i>MiSeq + HiSeq</i>	<i>MiSeq + HiSeq</i>	40%	60%	80%	40%	60%	80%
S115C41	9.426	1.236.206	1.245.632	494.004	493.996	227	825	2061	218	760	1856
S115B21	27.424	3.368.716	3.396.140	1.377.836	1.377.800	233	1196	3587	215	1065	3112
S115A8	4.714	544.756	549.470	221.312	221.306	168	443	999	162	404	889
S44V3	14.478	2.065.946	2.080.424	836.110	836.084	271	1096	3011	252	980	2598
S38V7B	30.682	3.977.334	4.008.016	1.637.056	1.636.987	306	1227	3662	287	1091	3155
S113F5	16.060	2.297.976	2.314.036	923.868	923.859	205	1124	2700	191	1009	2373
S113F7	15.004	2.053.212	2.068.216	826.914	826.888	310	1124	2904	289	1006	2510
S110F4	19.292	2.495.600	2.514.892	1.001.288	1.001.272	257	1218	3668	241	1089	3167
S77A7	7.554	849.228	856.782	347.272	347.271	251	769	1465	241	703	1333
S145F4	20.638	2.693.388	2.714.026	1.082.547	1.082.522	320	1149	2695	299	1034	2399
S149V11	19.056	2.230.784	2.249.840	921.437	921.421	255	1228	3182	240	1100	2793
S149V15	18.488	2.491.292	2.509.780	1.012.882	1.012.872	281	928	2091	266	852	1928
S152A12	16.796	2.195.350	2.212.146	888.801	888.791	223	508	1092	211	463	987
S152A4	40.310	4.312.916	4.353.226	1.786.124	1.786.090	261	1014	2851	245	901	2437
S77A30	37.014	4.206.910	4.243.924	1.745.447	1.745.442	321	1085	2926	303	974	2583
S131F4	36.718	3.575.122	3.611.840	1.492.376	1.492.331	209	846	2048	199	783	1882
S103D	33.564	3.305.236	3.338.800	1.389.894	1.389.878	321	1042	2526	301	942	2256
S99A6	27.670	2.832.994	2.860.664	1.194.718	1.194.709	186	743	1703	175	690	1583
S149V14	30.434	3.391.430	3.421.864	1.417.755	1.417.739	298	950	2519	279	848	2194
S77A31	30.940	3.473.546	3.504.486	1.446.493	1.446.488	243	990	2544	228	888	2282
S77A17	8.904	890.590	899.494	366.340	366.339	145	483	1104	138	444	1022
S38V6B	6.108	683.538	689.646	273.705	273.700	216	419	807	206	393	743
S151V3	13.704	1.333.428	1.347.132	559.474	559.451	152	399	919	148	382	886
S147F6	6.580	635.540	642.120	262.438	262.436	127	327	697	120	311	676

Tabela 2. Conclusão

Amostras	Nº leituras analisadas (não pareadas)			Nº leituras após o SeqClean (pareadas)	Nº leituras após o iPyRAD (pareadas)	Nº Loci (iPyRAD)			Nº Loci (CD-HIT)		
	<i>MiSeq</i>	<i>HiSeq</i>	<i>MiSeq + HiSeq</i>	<i>MiSeq + HiSeq</i>	<i>MiSeq + HiSeq</i>	40%	60%	80%	40%	60%	80%
S152A8	36.452	3.883.250	3.919.702	1.578.678	1.578.635	294	788	2088	275	702	1808
S137A1	30.816	3.139.166	3.169.982	1.289.755	1.289.734	282	916	2288	262	835	2093
S151V3	13.704	1.333.428	1.347.132	559.474	559.451	152	399	919	148	382	886
S150F1	23.208	2.475.290	2.498.498	1.028.418	1.028.410	276	569	1163	263	522	1077
S150F2	32.932	3.246.200	3.279.132	1.346.695	1.346.676	266	674	1734	250	621	1592
S149V16	18.660	1.991.324	2.009.984	813.791	813.773	226	595	1421	212	552	1316
S152A7	34.648	3.656.838	3.691.486	1.521.870	1.521.868	283	664	1724	264	597	1535
S152A15	39.082	4.077.608	4.116.690	1.729.014	1.728.929	194	480	1100	182	434	981
S152A13	39.044	4.106.192	4.145.236	1.714.294	1.714.208	176	625	1562	166	564	1382
S152A3	35.974	3.998.178	4.034.152	1.659.256	1.659.229	254	576	1321	240	527	1177
S152A2	25.878	3.061.426	3.087.304	1.247.702	1.247.683	246	534	1155	228	484	1044
S152A1	23.098	2.850.762	2.873.860	1.140.218	1.140.138	255	539	1163	240	490	1054
S26A16	22.240	2.861.860	2.884.100	1.129.187	1.129.185	186	520	1174	176	486	1110
S152A5	29.084	3.223.880	3.252.964	1.336.831	1.336.805	245	573	1334	227	516	1211
S35M1A	20.640	2.606.172	2.626.812	1.011.905	1.011.902	220	514	1071	204	467	989
S152A14	41.318	4.288.390	4.329.708	1.806.949	1.806.940	222	528	1297	209	481	1171
S62A21	-	6.269.966	6.269.966*	2.550.101*	2.550.047*	69	129	248	67	118	217

* apenas *HiSeq*

5.2 Análises Filogenéticas

A Tabela 3 sintetiza o tamanho da matriz de dados, em número de bases, gerada em cada cenário de triagem, sendo que para a matriz Loci, foram recuperados 333 (até 40%), 1440 (até 60%) e 6141 (até 80%) loci em cada um dos conjuntos de dados finais (após a retirada de sequências duplicadas identificadas no CD-HIT). Além disso, também pode-se observar o crescimento, de acordo com o cenário, do número de bases e sítios parcimoniosamente informativos, que são aqueles efetivamente utilizados para a reconstrução das relações de parentesco. A porcentagem de CG ficou bastante próxima daquela estimada para o genoma parcial de um indivíduo da subespécie *C. fernambucensis* (dados não publicados).

Tabela 3- Número de bases em cada matriz de dados, número de sítios informativos e constantes e porcentagem de CG para cada um dos cenários de dados faltantes.

Cenários	Nº bases por matriz de dados		Sítios		
	SNPs ligados	Loci	Parcimoniosamente informativos	Constante	% CG
40%	3694	70.186	1725	66.492	42
60%	21.442	334.256	10.885	312.814	40
80%	96.097	1.471.359	45.431	1.375.261	40

As figuras 4, 5 e 6 mostram matrizes de ocupação para cada um desses cenários e, além de observar o aumento no tamanho da matriz de dados conforme a tolerância à presença de dados faltantes é maior, também se observa a presença e ausência de cada loco em cada uma das amostras.

Já em relação aos diferentes programas usados para a reconstrução filogenética por Máxima Verossimilhança por meio de uma super matriz de dados concatenados, ambos (i.e. RaXML e IQTree), recuperaram topologias bastante semelhantes, sendo que a maioria das relações concordou com estudo publicado de Franco et al. (2017a), porém apresentando alto grau de suporte de relações antes recuperadas sem suporte e, também resolvendo relações terminais politômicas. A grande diferença entre esses dois programas é em relação ao tempo computacional para a realização das análises, onde o IQTree, como já esperado, realizou as análises em questão de horas, enquanto para o RaXML algumas análises se prolongaram por dias.

Figura 4 - Matriz de ocupação no cenário de até 40% de dados faltantes indicando presença e ausência dos loci em cada amostra (Tabela 1). Em vermelho estão destacadas amostras do *outgroup*. Os quadrados em preto indicam presença enquanto os brancos indicam ausência.

Figura 5 - Matriz de ocupação no cenário de até 60% de dados faltantes indicando presença e ausência dos loci em cada amostra (Tabela 1). Em vermelho estão destacadas amostras do *outgroup*. Os quadrados em preto indicam presença enquanto os brancos indicam ausência.

Figura 6 - Matriz de ocupação no cenário de até 80% de dados faltantes indicando presença e ausência dos loci em cada amostra (Tabela 1). Em vermelho estão destacadas amostras do *outgroup*. Os quadrados em preto indicam presença enquanto os brancos indicam ausência.

As árvores geradas para os três cenários de dados faltantes e tipos de informação (SNPs ligados e loci) estão apresentadas nos Apêndices de D a K. Apesar de bastante concordância entre as diferentes análises (Figura 7), o conjunto de dados com maior número de loci e sítios parcimoniosamente informativos (até 80% de dados faltantes; Tabela 3), foi considerado o mais informativo uma vez que possibilitou, além da recuperação das relações mais profundas e resolução de relações politômicas terminais (assim como o cenário de até 60% de dados faltantes) dentro do gênero *Cereus*, com valores de suportes significativos, também recuperou com mais definição as relações de *Cereus* com os gêneros *Cipocereus* e *Praecereus* (Figura 8; Figura 9). Assim, as relações evolutivas do gênero *Cereus* foram discutidas com base no maior conjunto de dados (Figura 8; Figura 9).

ANÁLISES DE MÁXIMA VEROSSIMILHANÇA		CLADO A			CLADO A1			CLADO A2			CLADO B		
		40%	60%	80%	40%	60%	80%	40%	60%	80%	40%	60%	80%
RaXML, <i>rapid bootstrap</i> , GTR + G	SNPs Ligados	100	100	100	51	97	100	57	97	100	100	100	100
	Loci	96	100	100	41	97	100	41	97	100	94	85	100
IQTree, <i>ultrafast bootstrap</i> , GTR + G	SNPs Ligados	100	100	100	54	100	100	57	100	100	100	100	100
	Loci	100	100	100	50	98	100	50	100	100	100	100	100
IQTree, SH- <i>alrt</i> , GTR + G	SNPs Ligados	100	100	100	99.5	100	100	100	100	100	100	100	100
	Loci	99.7	100	100	99.5	100	100	100	100	100	100	100	100
		CLADO C			CLADO D			CLADO E			CLADO F		
		40%	60%	80%	40%	60%	80%	40%	60%	80%	40%	60%	80%
RaXML, <i>rapid bootstrap</i> , GTR + G	SNPs Ligados	100	99	100	100	100	100	100	100	100	100	100	96
	Loci	100	100	100	94	90	98	100	100	100	100	100	100
IQTree, <i>ultrafast bootstrap</i> , GTR + G	SNPs Ligados	100	100	100	95	100	100	100	100	100	100	100	93
	Loci	100	100	100	95	100	100	100	100	100	100	100	100
IQTree, SH- <i>alrt</i> , GTR + G	SNPs Ligados	100	100	100	100	100	100	100	100	100	100	100	100
	Loci	100	100	100	99.9	100	100	100	100	100	100	100	100

Figura 7 - Esquema mostrando os suportes de ramos para os principais clados das filogenias geradas para *Cereus* neste trabalho (Figura 8; Figura 9, APÊNDICES de D a K), de acordo com os diferentes programas, medidas de suporte, matrizes de dados e cenários de dados faltantes. Onde, a cor verde indica suporte significativo e a cor amarela indica suporte não-significativo para cada uma das medidas de suporte de ramos (SH-aLRT >80, *ultrafast bootstrap* >95 e *rapid bootstrap* >70).

Figura 8 - Árvore filogenética colapsada estabelecida por Máxima Verossimilhança baseada na matriz SNPs ligados contendo 96.097 sítios e permitindo até 80% de dados faltantes gerada nos programas IQTree e RaXML. Os suportes de ramos foram considerados significativos quando maiores que 80 para Shimodaira-Hasegawa + aLRT (SH-aLRT), maiores que 95 para o *ultrafast bootstrap* (UFBOOT) e maiores que 70 para o *rapid bootstrap* (RBS), representados nessa ordem na árvore acima dos ramos. O * indica suporte de 100 para ambas as medidas (SH-aLRT, UFBOOT, RBS) e os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura. Os quadrados coloridos indicam as áreas de ocorrência das espécies do gênero *Cereus*.

Figura 9 - Árvore filogenética colapsada estabelecida por Máxima Verossimilhança baseada na matriz Loci contendo 6141 loci e permitindo até 80% de dados faltantes gerada nos programas IQTree e RaXML. Os suportes de ramos foram considerados significativos quando maiores que 80 para Shimodaira-Hasegawa + aLRT (SH-aLRT), maiores que 95 para o *ultrafast bootstrap* (UFBOOT) e maiores que 70 para o *rapid bootstrap* (RBS), representados nessa ordem na árvore acima dos ramos. O * indica suporte de 100 para ambas as medidas (SH-aLRT, UFBOOT, RBS) e os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura. Os quadrados coloridos indicam as áreas de ocorrência das espécies do gênero *Cereus*.

Dessa forma, os principais resultados filogenéticos obtidos foram:

- 1) O gênero *Cereus* é composto por quatro linhagens principais (A-D) e pode ser monofilético, uma vez que as espécies dos gêneros *Cipocereus* (clado E) e *Praecereus* (clado F) foram excluídas do *ingroup*, apesar de suporte significativo (SH-aLRT >80, UFBOOT >95 e RBS >70) para a exclusão de *Cipocereus* ter sido recuperado apenas para a topologia gerada a partir da matriz de SNPs ligados no RaXML (Figura 8; Figura 9);
- 2) Os subgêneros *Cereus* e *Ebneria* não formaram grupos monofiléticos em contraste com o subgênero *Mirabella* que foi resolvido em um grupo coeso (Figura 8; Figura 9). Como há apenas um representante do subgênero *Oblongicarpus* não foi possível testar a monofilia desse subgênero;
- 3) As espécies do subgênero *Mirabella* formam um grupo monofilético (clado C), mas possivelmente não relacionado à *Cipocereus*, como sugerido anteriormente (FRANCO et al. 2017a);
- 4) As espécies do clado A formam um grupo monofilético com as espécies do clado B;
- 5) As espécies do clado A foram subdivididas em dois subclados (A1 e A2), concordando com resultados prévios (Figura 2). O posicionamento intermediário entre os clados A1 e A2 da amostra de *C. jamacaru jamacaru* (S110F4), proveniente de uma região de restinga no nordeste brasileiro, sugere que houve hibridação após o contato secundário de *C. jamacaru* e *C. fernambucensis* no litoral;
- 6) As relações evolutivas encontradas no subclado A2 indicam a subdivisão da espécie *C. fernambucensis* em grupos populacionais e a parafilia da subespécie *C. fernambucensis fernambucensis*, corroborando Franco et al. (2017b);
- 7) *C. saddianus* e *C. kroenleinii* formam um grupo monofilético (clado D), mas não relacionado à *Praecereus* como sugerido anteriormente (FRANCO et al. 2017a).

6. DISCUSSÃO

6.1 Considerações a respeito de diferentes protocolos ddRADSeq

Com base nos resultados insatisfatórios da Biblioteca I (APÊNDICE A - Tabela 1), que impossibilitariam a continuidade do trabalho e alcance dos objetivos, buscou-se por problemas que possivelmente teriam acontecido nas etapas deste protocolo. Algumas etapas durante a preparação de uma biblioteca reduzida são consideradas fundamentais para o sucesso do sequenciamento, entre essas etapas a obtenção do DNA genômico é a primeira com a qual qualquer interessado em metodologias de redução do genoma depara-se. A qualidade e a concentração do DNA alvo durante a digestão podem comprometer o sucesso do trabalho (MCCOMARCK et al. 2013; PURITZ et al. 2014), como por exemplo, a baixa cobertura durante o sequenciamento observada quando amostras com DNA degradado são utilizadas (GRAHAM et al. 2015). Ao compararmos as Bibliotecas I (APÊNDICE A) e II, pode-se observar que ao utilizar maior concentração de DNA, melhores são os resultados do sequenciamento. Apesar do que foi defendido por Kess et al. (2016) e Yang et al. (2016), aqui percebemos que para protocolos ddRadSeq uma grande quantidade de DNA genômico (a partir de 200 ng) parece ser importante para o sucesso do sequenciamento, o que também foi observado em bibliotecas reduzidas para espécies dos gêneros, também de Cactaceae, *Melocactus* e *Uebelmannia* (informação pessoal)¹.

Outra etapa importante é a ligação de adaptadores, sendo a super-representação de sequências similares aos adaptadores Illumina no sequenciamento um dos problemas observados na Biblioteca I (APÊNDICE A). Nessa biblioteca, após as filtragens no SeqyClean e no PyRAD, foi perdido entorno de 50% dos dados, sendo que no final obtivemos aproximadamente 10% de aproveitamento do sequenciamento (APÊNDICE B - Tabela 1). Etapas de seleção de tamanho/purificação após a ligação e antes do enriquecimento por PCR podem ser a solução para desvios na etapa da ligação. De fato, Yang et al. (2016) desenvolveram outros dois protocolos para a montagem de bibliotecas modificados de Peterson et al. (2012), onde no protocolo A houve seleção de tamanho/purificação após a PCR e no protocolo B a seleção de tamanho/purificação ocorreu antes da PCR. No protocolo A cerca de 50% de contaminação por adaptadores foi observada, enquanto no protocolo B apenas 3% das sequências foram perdidas por excesso de adaptador.

¹ Comunicação entre integrantes (2017) do Laboratório de Diversidade Genética e Evolução (LaGEvol – Departamento de Biologia – UFSCar, campus Sorocaba).

O protocolo de Campos et al. 2017 utilizado na Biblioteca II, realiza passos de purificação com *beads* e quantificação após cada etapa da biblioteca reduzida, além da seleção de tamanho realizada antes do enriquecimento por PCR, e esses passos provavelmente evitaram excesso de adaptadores não ligados que poderiam ter sido amplificados na PCR, como aconteceu na Biblioteca I. Essas etapas de purificação após a ligação dos adaptadores e a seleção de tamanho antes da amplificação por PCR no protocolo da Biblioteca II, ocasionaram a diminuição, aproximadamente, de 50 % para 0,5-2% do excesso de adaptadores filtrados no SeqClean.

Além da necessidade de DNA genômico em grande quantidade e de boa qualidade e do excesso de adaptadores, na literatura (MCCOMARCK et al. 2013; PURITZ et al. 2014) a maior desvantagem destes protocolos é a seleção de tamanho manual, o que pode ser resolvido com seleção automática, como por exemplo, o Pippin Prep (Sage Science). Neste trabalho, fizemos um teste de seleção de tamanho do *pool* da Biblioteca II no Pippin Prep, porém houve uma perda considerável da concentração do *pool* (0,26 nM após o Pippin Prep) e como a amplitude dos picos no *Bioanalyzer* foi semelhante, na seleção manual e na automática, optamos pela manual.

Um obstáculo presente em ambas as bibliotecas, é a baixa complexidade (pouca variação entre as sequências de DNA dos indivíduos) entre amostras nos primeiros ciclos do sequenciamento, o que é resultado de uma característica intrínseca na montagem da biblioteca desses protocolos ddRADSeq, onde a leitura das sequencias começa pelos sítios de restrição, que é idêntico para todos os indivíduos e atrapalha a identificação de grupos (“*clusters*”) pelo sequenciador, o que nos levou a uma perda de cerca de 40% dos dados na Biblioteca I. Para contornar este problema no sequenciamento da Biblioteca II optamos por inserir o *PhiX*, que é uma biblioteca de um bacteriófago com sequência conhecida e padronizada pela Illumina, que objetiva aumentar a complexidade de bibliotecas, assim aproximadamente 18% do rendimento deste sequenciamento foi consumido por *PhiX*, um valor menor do que o perdido por baixa complexidade no sequenciamento da Biblioteca I.

Diante dessas considerações, é sugerida uma avaliação prévia cuidadosa dos protocolos de bibliotecas reduzidas de acordo com os objetivos pretendidos e os recursos biológicos disponíveis. Além disso, simulações *in silico* podem auxiliar em tomadas de decisão importantes durante o processo de preparação da biblioteca. Outra consideração importante é sobre o controle de qualidade da biblioteca em relação a quantificação, espectro de tamanho selecionado e indicativos de excesso de adaptadores, já que para a Biblioteca I a amplitude da

curva do *Bioanalyzer* foi baixa e a concentração (qPCR) foi mais alta, ao contrário da Biblioteca II, indicando o *Bioanalyzer* como um ponto de checagem de qualidade importante durante o desenvolvimento da biblioteca.

6.2 Análise filogenéticas

Existe atualmente muitos conflitos de opiniões com relação a inclusão ou exclusão de dados faltantes no conjunto final de dados para análises filogenômicos (HUANG; KNOWLES, 2014; LEACHÉ et al. 2015; HOSNER et al. 2015). O aumento da matriz é a primeira evidência ao permitir a inclusão de dados faltantes no conjunto final de loci, sendo este aumento atraente para estudos que buscam a resolução de relações de parentesco em grupos de radiação recente e explosiva (HUANG; KNOWLES, 2014; STREICHER; SCHULTE; WIENS, 2016), porém a presença de dados faltantes pode gerar discordância topológica, erros na estimação de comprimento de ramos e decréscimo nos valores de *bootstrap* (LEACHÉ et al. 2015).

A presença de dados faltantes, até o momento, é praticamente inevitável nas tecnologias de sequenciamento em massa, porém a proporção com que esses dados aparecem envolvem desde questões biológicas até processamento de dados. Durante a preparação de bibliotecas RADSeq processos biológicos, como mutações nos sítios de corte das enzimas podem gerar alelos nulos, fazendo com que determinadas regiões mais variáveis não sejam recuperadas para alguns indivíduos, o que é proporcional à distância genética da amostragem taxonômica incluída na biblioteca. Ao observarmos as Figuras 4, 5 e 6, percebe-se que com menos dados faltantes as amostras do *outgroup* encontram-se mais espalhadas na matriz, o que pode ser reflexo da permanência de loci pouco variáveis, porém conforme a presença de dados faltantes é maior, essas amostras concentram-se na parte inferior da matriz, onde o número de dados faltantes é maior, indicando a manutenção de loci mais variáveis, que foram aqueles capazes de resolver com mais clareza as relações filogenéticas de *Cereus*.

Outro ponto é a estocasticidade durante o sequenciamento de milhares de loci, podendo gerar número variável de leituras entre as amostras durante o sequenciamento. Durante o processamento, as decisões como similaridade entre as sequências e o número mínimo de amostras em que o loco tem que estar presente para ser mantido, afetam diretamente a quantidade de dados faltantes no conjunto final de loci (MCCOMARCK et al. 2013; HUANG; KNOWLES, 2014; HOSNER et al. 2015).

No presente estudo parece que a presença de dados faltantes está mais ligada às questões biológicas e à aleatoriedade do sequenciamento. Como pode ser observado na Tabela 2, para a grande maioria das amostras o número de leituras se manteve entorno de dois e três milhões, mas em alguns casos como para S76F10 (10.259.434) e S147F6 (642.120), houveram desvios no rendimento do sequenciamento. Esse fato também pode estar relacionado com a qualidade da amostra de DNA genômico utilizado na preparação da biblioteca, onde amostras degradadas podem gerar baixa cobertura durante o sequenciamento (GRAHAM et al. 2015).

O uso de tecnologias NGS para obtenção de dados e sua aplicação em filogenética molecular, ainda é recente, o que gera falta de consenso em como aplicar esses dados da melhor maneira possível (HUANG; KNOWLES, 2014; DUPUIS et al. 2017). Uma visão mais conservadora pode optar por excluir dados faltantes de análises filogenéticas e filogeográficas (ex.: MCCOMARCK et al. 2013) defendendo que análises populacionais, filogeográficas e coalescentes são sensíveis à presença de dados faltantes, além da alta exigência computacional imposta por grandes matrizes de dados em análises filogenéticas e também pelo fato de alguns trabalhos indicarem que conjuntos de dados com altos níveis de dados faltantes podem estar mais susceptíveis a erros (EDWARDS, 2008; HOSNER et al. 2015; LEACHÉ et al. 2015).

Neste trabalho, diferente do observado por Leaché et al. (2015), uma análise visual das topologias apresentadas nos permite apontar que não houve decréscimo nos valores de *bootstrap* conforme o volume de dados faltantes foi maior, pelo contrário, de certa maneira esses valores foram bem semelhantes para algumas relações e, que aparentemente, se comparadas as árvores de todos os cenários de dados faltantes não houve discordância topológica e sim um aumento da resolução filogenética, como observado por outros autores (HUANG; KNOWLES, 2014).

Em relação à exigência computacional, como mostrado aqui, novos programas estão sendo desenvolvidos objetivando a análise de grandes conjuntos de dados, resultando em árvores robustas e confiáveis quando comparadas com clássicos programas. Essa questão também pode ser resolvida ao se trabalhar com matrizes apenas com sítios variáveis, o que reduz consideravelmente o volume de dados (Tabela 3) e oferece resultados semelhantes a matrizes maiores contendo sítios variáveis e invariáveis (Figura 7; Figura 8; Figura 9), sempre considerando o uso de correções capazes de diminuir os efeitos de comprimento de ramo quando utilizado apenas sítios variáveis. Leaché et al. (2015) observou que alinhamentos contendo apenas sítios variáveis empregados em análises filogenéticas sem correção,

superestimam em mais de quatro vezes o comprimento da árvore e que esses mesmos alinhamentos empregados com correções fornecem comprimentos semelhantes aos alinhamentos contendo sítios invariantes quando existem poucos dados faltantes, sendo que o comprimento da árvore foi levemente superestimado conforme o aumento da presença de dados faltantes.

O maior impacto do aumento da matriz de dados gerado pela alta permissibilidade a dados faltantes neste trabalho, foi a maior resolução filogenética e aumento nos suportes de ramos, o que nos permitiu inferir a monofilia do gênero *Cereus*. Ao observar as topologias geradas a partir das matrizes de até 40% de dados faltantes (APÊNDICES de D a G), apesar dessas já resolverem as relações terminais politômicas da Figura 2, percebe-se fraca resolução filogenética para algumas linhagens, uma vez que relações mais profundas foram pouco suportadas, o que de maneira crescente foi melhor recuperado nas topologias geradas a partir do cenário de até 60% (APÊNDICES H a K) e até 80% de dados faltantes (Figura 8; Figura 9).

Resultados semelhantes em relação a fraca resolução quando níveis menores de dados faltantes são permitidos veem sendo encontrados em vários grupos, como por exemplo em mosquitos (EMERSON et al. 2010), ciclídeos (WAGNER et al. 2013), bactérias (PETTENGIL et al. 2014) e lagartos (STREICHER; SCHULTE; WIENS, 2016). Huang e Knowles (2014), a partir de dados RADSeq simulados, observaram o aumento da precisão filogenética com o aumento de dados faltantes, especialmente para espécies de radiação recente, onde uma melhor resolução sobre o histórico de divergência dos táxons pode depender do tamanho da matriz de dados. Wagner et al. (2013) mostrou o poder da inclusão de um grande volume de dados faltantes para resolver as relações de parentesco entre espécies de ciclídeos, onde sua melhor topologia foi gerada a partir de uma matriz com mais de 90% de dados faltantes, o que segundo ele permitiu a resolução de questões antes nunca resolvidas para este grupo diverso e de origem recente (< 15000 anos).

A inclusão de dados faltantes em análises filogenéticas permite com que loci mais variáveis sejam mantidos, assim a simples exclusão de todos os dados faltantes pode ser problemática, onde topologias geradas a partir de um conjunto de loci pouco permissivo a presença de dados faltantes, muitas vezes são menos precisas do que as estimadas a partir de conjuntos de dados mais permissivos (HUANG; KNOWLES, 2014; STREICHER; SCHULTE; WIENS, 2016; REDDY et al. 2017). Porém, apesar de quase sempre inclusão de mais loci, pelo aumento de dados faltantes, aumentar o suporte de ramo, não existe até o momentos regras

gerais para seu uso (HUANG; KNOWLES, 2014; LEACHÉ et al. 2015), assim sugere-se o uso desses dados com cuidado, avaliando a importância do aumento da matriz no contexto do grupo estudado e fazendo, sempre que possível, um aumento gradual de dados faltantes no conjunto de dados, observando a quantidade de dados necessária pra melhores resultados em relação a precisão filogenética, uma vez que conjunto de dados com mais dados faltantes podem estar sujeitos a erros (PETTENGIL et al. 2014; LEACHÉ et al. 2015) e nem sempre o aumento infinito da matriz irá melhorar a resolução filogenética (ZELLMER et al. 2012), assim é importante um equilíbrio entre número de loci/SNPs e porcentagem de dados faltantes, o que dependerá principalmente do número de táxons amostrados e a distância genética entre eles.

Esse aumento substancial da quantidade de dados/marcadores tem muito a contribuir com a resolução de perguntas na biologia evolutiva, porém os programas para a análise filogenética desses dados precisam estar preparados para matrizes de dados filogenômicos, uma vez que o volume e a complexidade dos dados estão cada vez maiores.

O RaXML é um programa de verossimilhança clássico e robusto, porém utiliza um algoritmo de rearranjo topológico que realiza uma busca heurística extensiva por todo espaço amostral de árvores, o SPR, sendo extremamente lento (ZHOU et al. 2017), o que dificulta seu uso em matrizes de dados filogenômicos. Em contraposição, o IQTree, um recente programa desenvolvido a partir da modificação do programa Tree Puzzle, tem objetivo de atender à dados filogenômicos e vem sendo cada vez mais utilizado. O NNI é algoritmo de rearranjo topológico usado pelo IQTree, sendo ele mais rápido que o SPR (ZHOU et al. 2017), o que o torna um programa interessante do ponto de vista de exigência computacional. No entanto, apenas uma economia em relação ao tempo da análise não seria o suficiente para a disseminação de seu uso. Trabalhos recentes (REDDY et al. 2017; ZHOU et al. 2017) estão buscando por comparações entre esses programas e o IQTree tem mostrado resultados iguais ou melhores que o RaXML, tanto em relação a exigência computacional como em relação aos suportes de ramo e resolução de relações de parentesco. Zhou et al. (2017) testou a eficiência de vários programas para geração de árvores de Máxima Verossimilhança e destacou a eficiência do IQTree em relação aos outros programas na análise de dados concatenados em uma supermatriz, porém também verificou a queda na eficiência deste programa quando aplicado a conjuntos de dados com mais de 200 táxons.

Os dados apresentados aqui ressaltam a similaridade entre os resultados gerados nas duas abordagens de verossimilhança, ressaltando a acurácia (Figuras 8; Figura 9) e a precisão

(Figura 7) dos resultados obtidos no IQTree, sendo esta uma importante observação uma vez que é esperado que os valores do UFBOOT implementado no IQTree sejam maiores em relação ao RBS do RaXML, pois o algoritmo de *bootstrap* utilizado pelo primeiro é mais sensível a possíveis violações de modelos, gerando superestimação dos valores de suporte de ramos (REDDY et al. 2017).

A partir disso, observou-se que todas as topologias geradas no cenário de até 80% de dados faltantes recuperaram a monofilia do gênero *Cereus* (Figura 8; Figura 9). Em relação a *Praecereus* (clado F), todas as topologias o mostram como um ramo fortemente suportado, diferenciado e monofilético, mesmo com a grande distância geográfica que separa os indivíduos amostrados de duas espécies representativas do gênero. Esse gênero já foi, junto com *Ebneria*, agrupado por alguns autores com base em caracteres morfológicos, em um gênero conhecido como *Monvillea* - Britton & Rose – que mais tarde foram transferidos para *Cereus* fazendo parte do subgênero *Ebneria* e atualmente é considerado um gênero distinto, o que é corroborado pela hipótese filogenética aqui apresentada (Figura 8; Figura 9) (HUNT, 1988; HUNT; TAYLOR; CHARLES, 2006; FRANCO et al. 2017a). Já para *Cipocereus* (clado E), o suporte de ramo significativo foi recuperado apenas para a matriz de SNPs ligados no RaXML (Figura 8), o que pode ter sido resultado de uma atração de ramos desse clado com o clado do subgênero *Mirabella* (clado C) (Figura 8; Figura 9) ou até por um agrupamento espúrio, considerando os efeitos que dados faltantes e alinhamentos apenas com sítios variáveis podem causar (LEACHÉ et al. 2015).

O gênero *Cipocereus*, endêmico de regiões de Cerrado em Minas Gerais (TAYLOR; ZAPPI, 2004), já foi taxonomicamente classificado como *Cereus* (pela semelhança de caracteres florais) e, também, como *Pilosocereus* (pela semelhança de caracteres florais, caules e frutos). Importante salientar que um recente estudo que investigou as relações de parentesco no gênero *Pilosocereus*, com múltiplos marcadores moleculares (ROMEIRO-BRITO, 2017), recuperou sua inclusão dentro deste gênero. Contudo, a sistemática de *Cipocereus* parece ainda ser confusa e necessita de mais estudos que possam esclarecer de fato sua relação com outros gêneros próximo, assim o resultado aqui apresentado deve ser analisado com ressalvas.

O clado A apresentado nas Figuras 8 e 9 com alto suporte de ramo, é o mesmo recuperado na filogenia anteriormente proposta para o gênero *Cereus* (Figura 2), o que mostra a robustez deste agrupamento, porém com maior número de marcadores pudemos resolver as relações politômicas das espécies terminais. Esse clado é composto apenas por espécies do

subgênero *Cereus* e tem por característica abrigar espécies que ocorrem, majoritariamente, na Caatinga e no Cerrado, representadas no subclado A1, com exceção de *C. hildmannianus*, que assim como as espécies do subclado A2, também ocorre na Mata Atlântica (e também no Pampa), porém em ambientes distintos (Figura 8; Figura 9). Esses dados podem corroborar com as hipóteses biogeográficas de Franco et al. (2017a), que sugerem um possível evento vicariante ocorrido entre a Mata Atlântica e o Cerrado, com expansão de florestas úmidas, durante o Pleistoceno, porém isso só poderá ser confirmado após serem realizadas datações e reconstruções biogeográficas a partir dessa nova hipótese filogenética.

O subclado A2, formado pelas espécies *C. fernambucensis* e *C. insularis*, é interessante para estudos populacionais e biogeográficos por ocorrerem em ambientes xéricos da Mata Atlântica, incluindo vegetações costeiras e afloramentos rochosos e, apesar de um estudo revelar a presença de 4 grupos genéticos entre essas espécies, o relacionamento filogenético entre eles não foi resolvido (FRANCO et al. 2017b), sendo assim os dados filogenômicos gerados e as relações de parentesco mais resolvidas apresentadas neste trabalho tem muito a contribuir com questões populacionais dessas espécies. Além disso, tanto *C. fernambucensis* quanto *C. insularis* são espécies endêmicas da Mata Atlântica, porém a primeira ocorre ao longo da região costeira do Brasil e a segunda é encontrada a 350 Km do litoral brasileiro no oceano Atlântico, na Ilha Fernando de Noronha (HUNT; TAYLOR; CHARLES, 2006), o que nos coloca frente à questões conservacionistas, uma vez que as regiões de Mata Atlântica no leste do Brasil são fortemente ameaçadas pela agricultura/aquicultura, superexploração de recursos e desenvolvimento humano (GOETTSCHE, et al. 2015) e uma melhor caracterização dessas espécies a partir desse grande conjunto de dados tende a contribuir para sua manutenção no ambiente natural.

Os clados B e D (Figura 8; Figura 9), são respectivamente os clados B e C1 recuperados por Franco et al. (2017a), sendo compostos pelas espécies dos subgêneros *Cereus* (*C. cochabambensis*, *C. hankeanus* e *C. stenogonus*) e *Ebneria* (*C. aethiops*, *C. kroenleinii*, *C. saddianus* e *C. spegazzinni*). As espécies do subgênero *Cereus* que posicionam-se o clado B, o que faz com que esse subgênero não seja monofilético, tem por característica estrutura colunar e são morfologicamente semelhantes as espécies do subgênero *Cereus* encontradas no clado A, com o qual forma um grupo monofilético, porém a área de ocorrência das espécies do clado B é o Chaco (HUNT; TAYLOR; CHARLES, 2006), assim como suas espécies irmãs do subgênero *Ebneria*, o que pode justificar a estreita relação filogenética entre essas espécies. Em

relação ao subgênero *Ebneria*, enquanto os indivíduos amostrados de *C. spegazzinni* e *C. aethiops* ocorrem no Chaco, os de *C. kroenleinii* e *C. saddianus* ocorrem em regiões pantaneiras, sendo que essa ocorrência em diferentes biomas pode explicar seu posicionamento, mesmo próximo, em clados diferentes.

Na filogenia anteriormente proposta para o gênero *Cereus* (Figura 2) é possível observar a falta de resolução nas relações dos ramos terminais e, também, os baixos valores de suporte para relações mais internas, sendo considerados por Franco et al. (2017a) como ruídos da radiação rápida e recente dentro do grupo. A hipótese filogenética proposta aqui (Figura 8; Figura 9) foi capaz de recuperar com clareza e alto suporte as relações das espécies do gênero *Cereus*, resolvendo as relações politômicas dos terminais e, também, relações mais profundas. Essas observações nos permitem, mesmo sem termos realizado a datação para essa nova hipótese filogenética, discutir a hipótese de rápida e recente radiação dentro do gênero *Cereus*.

Além da rápida radiação, grupos que sofreram diversificação recente podem não ter acumulado diferenças suficientes entre linhagens distintas, onde um maior número de marcadores moleculares tende a fornecer mais sinal filogenético, acarretando em relações de parentesco bem estabelecidas, o que foi observado neste trabalho, contribuindo com hipóteses de que o aumento da variação genética tem potencial para resolver relações de parentesco, antes mal compreendidas, em cenários de radiação rápida e recente.

Mesmo para os ramos internos, aparentemente curtos com nós próximos, nosso conjunto de dados foi capaz de esclarecer, na maioria das vezes, essas relações mais antigas que pelo comprimento dos ramos provavelmente foram formadas em um evento de diversificação explosiva em um intervalo de tempo menor do que as linhagens mais novas, o que dificulta ainda mais o estabelecimento dessas relações filogenéticas e, muitas vezes, mesmo com um grande volume de dados não são resolvidas (GIARLA; ESSELSTYN, 2015; WELCH et al., 2016).

7. CONCLUSÕES

O uso do *NGS* apresentou-se como uma ótima alternativa na obtenção de um conjunto de dados genômico e com grande custo-benefício se comparado a outras metodologias (ex: sequenciamento de Sanger). O protocolo ddRADSeq gerou um subconjunto de regiões do genoma que proporcionou maior profundidade de cobertura por loco e o sequenciamento de mais amostras, levando a obtenção de um grande volume de dados filogenômicos que foi capaz de resolver questões filogenéticas do gênero *Cereus* antes não resolvidas por baixo sinal filogenético. Além disso, a padronização de um protocolo de biblioteca reduzida, que além do gênero *Cereus* já foi aplicado com sucesso em dois outros gêneros de cactos (*Melocactus* e *Uebelmannia*) deverá contribuir para muitos outros trabalhos de Cactaceae.

A permissividade à presença de dados faltantes foi essencial para a recuperação de relações filogenéticas bem definidas neste trabalho, porém o uso dessa abordagem não deve ser indiscriminado e, sim norteado por necessidades impostas pelo grupo de estudo e sempre buscando por uma análise de dados cuidadosa. Como consequência de manter dados faltantes têm-se o aumento da matriz de dados, o que junto com o crescente aumento na geração de dados fazem com que seja de fundamental importância investigar a relação de desempenho de diferentes programas de análises filogenéticas, uma vez que recursos computacionais ainda são escassos. Porém, esses programas devem apresentar condições vantajosas não somente em relação ao tempo computacional, eles devem também gerar resultados eficientes e precisos quando comparados a programas que empregam algoritmos mais lentos.

Como esperado, o suporte e a resolução das relações de parentesco do gênero *Cereus* aumentaram significativamente com o aumento de mais sítios informativos, o que pode colaborar para uma revisão taxonômica desse grupo amplamente distribuído na América do Sul, mas que ainda sofre com incertezas taxonômicas. O nosso conjunto de dados possui sinal suficiente para fornecer informações sobre as relações internas e, principalmente, sobre as relações terminais do gênero, o que nos permitiu inferir a monofilia de *Cereus*, confirmar a monofilia do subgênero *Mirabella*, a formação de um grupo monofilético coeso formado pelos clados A e B e a parafilia da espécie *C. fernambucenis*.

REFERÊNCIAS

- ALLENDORF, F. W. Genetics and the conservation of natural populations: allozymes to genomes. **Molecular Ecology**, v. 26, p. 420-430, Nov. 2017.
- ANDREWS, K. R. et al. Trade-offs and utility of alternative RADseq methods. **Molecular Ecology** v. 23, p. 5943-5946, Dez. 2014.
- ANDREWS, K. R. et al. Harnessing the power of RADseq for ecological and evolutionary genomics. **Nature Reviews Genetics**, v. 17. n. 2, p. 81-92, Feb. 2016.
- ANISIMOVA, M; GASCUEL, O. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. **Systematic Biology**, v. 55, n. 4, p. 539–552, Ago. 2006.
- ARAKAKI, M. et al. Contemporaneous and recent radiations of the world's major succulent plant lineages. **Proceedings of the National Academy of Sciences**, v. 108, n. 20, p. 8379-8384, Mai. 2011.
- AVISE, J. C. Phylogeography: retrospect and prospect. **Journal of biogeography**, v. 36, n. 1, p. 3-15, Nov. 2009.
- AVISE, J. C; BOWEN, B. W; AYALA, F. J. In the light of evolution X: comparative phylogeography. **PNAS**, v. 113, n. 29, p. 7957–7961. Jul. 2016.
- BARTHLOTT, W. et al. *Biogeography and Biodiversity of Cacti*. Schumannia, Jan. 2015.
- BELFIORE, N. M.; LIU, L; MORITZ, C. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). **Systematic Biology**, v. 57, n. 3, p. 294-310, Jun. 2008.
- BONATELLI, I. A.S. et al. Interglacial microrefugia and diversification of a cactus species complex: phylogeography and palaeodistributional reconstructions for *Pilosocereus aurisetus* and allies. **Molecular ecology**, v. 23, n. 12, p. 3044-3063, Abr. 2014.
- BONATELLI, I. A. S; CARSTENS, B. C; MORAES, E. M. Using Next Generation RAD Sequencing to Isolate Multispecies Microsatellites for *Pilosocereus* (Cactaceae). **PloS one**, v. 10, n. 11, p. 1-14, Nov. 2015.

- BRITO, P. H; EDWARDS, S. V. Multilocus phylogeography and phylogenetics using sequence-based markers. **Genetica**, v. 135, n. 3, p. 439-455, 2009.
- BOROWIEC, M. L. AMAS: a fast tool for alignment manipulation and computing of summary statistics. **PeerJ**, v. 28, n. 4, Jan. 2016.
- BUTTERWORTH, C. A; WALLACE, R. S. Phylogenetic studies of *Mammillaria* (Cactaceae)—insights from chloroplast sequence variation and hypothesis testing using the parametric bootstrap. **American Journal of Botany**, v. 91, n. 7, p. 1086–1098, Fev. 2004.
- CALVENTE, A. et al. Molecular phylogeny of tribe *Rhipsalideae* (Cactaceae) and taxonomic implications for Schlumbergera and Hatiora. **Molecular Phylogenetics and Evolution**, v. 58, n. 3, p. 456-468, 2011.
- CALVENTE, A. et al. Phylogenetic analyses of *Pilosocereus* (Cactaceae) inferred from plastid and nuclear sequences. **Botanical Journal of the Linnean Society**, v. 183, n. 1, p. 25-38, 2017.
- DACOSTA, J. M; SORENSON, M. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. **Molecular Phylogenetics and Evolution**, v. 94, p. 122-135, Jul. 2015.
- DAVEY, J. W; BLAXTER, M. L. RADSeq: next-generation population genetics. **Brief Funct Genomics**, v. 9, n. 5, Dez. 2010.
- DIJK, E. L. et al. Ten years of next-generation sequencing technology. **Trends Genetic**, v. 30, n. 9, p. 418-26, Ago. 2014.
- DOYLE, J. J. The promise of genomics for a “next generation” of advances in higher-level legume molecular systematics. **South African Journal of Botany**, v. 89, p.10-18, Jul. 2013.
- DUPUIS, J. R. et al. Genome-wide SNPs resolve phylogenetic relationships in the North American spruce budworm (*Choristoneura fumiferana*) species complex. **Molecular Phylogenetics and Evolution**, v. 111, p. 158–168, Abr. 2017.
- EATON, D. A. R. et al. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. **Bioinformatics**, v. 30, n. 13, p. 1844–1849, Jul. 2014.
- EDWARDS, S. V. Is a new and general theory of molecular systematics emerging? **Evolution**, v. 63, n. 1, p. 1–19, Out. 2008.

- ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS One**, v. 6, n. 5, Mai. 2011.
- EMERSON, K. J. et al. Resolving postglacial phylogeography using high-throughput sequencing. **PNAS**, v. 107, n. 37, p. 16196–16200, Set. 2010.
- FENG, Y. et al. Nanopore-based Fourth-generation DNA Sequencing Technology. **Genomics, Proteomics & Bioinformatics**, v. 13, n. 1, p. 4-16, Fev. 2015.
- FRANCO, F. F. et al. Plio-Pleistocene diversification of *Cereus* (Cactaceae, Cereeae) and closely allied genera. **Botanical Journal of the Linnean Society**, v. 183, n. 2, p. 199-210, Out. 2017a.
- FRANCO, F. F. et al. The xeric side of the Brazilian Atlantic Forest: The forces shaping phylogeographic structure of cacti. *Ecology and Evolution*, v. 7, p. 9281-9293, Ago. 2017b.
- FRANCK, A. R.; COCHRANE, B. J.; GAREY, J. R. Phylogeny, biogeography, and infrageneric classification of *Harrisia* (Cactaceae). **Systematic Botany**, v. 38, n. 1, p. 210-223, 2013.
- FRANCK, A. R. et al. A new species of *Cereus* S.STR. (Cactaceae) endemic to Haiti. **Phytoneuron**, v. 29, p. 1–17, Abr. 2017.
- FU, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics Applications Note**, v. 28, n. 23, p. 3150–3152, Out. 2012.
- GIARLA, T. C; ESSELSTYN, J. A. The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews. **Systematic Biology**, v. 64, n. 5, p. 727–740, Mai. 2015.
- GOETTSCHE, B. et al. High proportion of cactus species threatened with extinction. **Nature plants**, v. 1, p. 15142, Out. 2015.
- GRAHAM, C. F. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). **Molecular Ecology Resources**, v. 15, p. 1304–1315, Abr. 2015.
- GUINDON, S. et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. **Systematic Biology**, v. 59, n. 3, p. 307–321, Mai. 2010.

HARVEY, M. G. et al. Sequence Capture Versus Restriction Site Associated DNA Sequencing for Phylogeography. **Systematic Biology**, v. 65, p. 910-924, Dec. 2013.

HERNÁNDEZ-HERNÁNDEZ, T. et al. Phylogenetic relationships and evolution of growth form in Cactaceae (Caryophyllales, Eudicotyledoneae). **American Journal of Botany**, v. 98, n. 1, p. 44-61, Dez. 2011.

HERNÁNDEZ-HERNÁNDEZ, T. et al. Beyond aridification: multiple explanations for the elevated diversification of cacti in the New World Succulent Biome. **New phytologist**, v. 202, n. 4, p. 1382-1397, Feb. 2014.

HOANG, D. T. et al. UFBoot2: Improving the Ultrafast Bootstrap Approximation. **Molecular Biology Evolution**, v. 35, n. 2, p. 518–522, Out. 2017

HOSNER, P. A. et al. Avoiding Missing Data Biases in Phylogenomic Inference: Na Empirical Study in the Landfowl (Aves: Galliformes). **Molecular Biology Evolution**, v. 33, n. 4, p. 110-1125, Dez. 2015.

HUANG, H; KNOWLES, L. L. Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. **Systematic Biology**, v. 65, n. 3, p. 357-365, Jun. 2014.

HUNT, D. New and unfamiliar names for use in the European Garden Flora: addenda and corrigenda. **Bradleya**, v. 6, p. 100, Dez. 1988.

HUNT, D; TAYLOR, N.; CHARLES, G. *The new cactus lexicon*. DH Books, Milborne Port, UK, 2006.

LEACHÉ, A. D. et al. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. **Systematic Biology**, v. 64, n. 6, p. 1032–1047, Jul. 2015.

LELIAERT, F. et al. DNA-based species delimitation in algae. **European Journal of Phycology**, v. 49, n.2, p.179-196, Mai. 2014.

LEPAIS, O; WEIR, J. T. SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. **Molecular Ecology Resource**, v. 14, n. 6, p. 1314-21, Nov. 2014.

- LI, W; GODZIK A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, v. 22, n. 13, Jul. 2006.
- MADDISON, Wayne P. Gene trees in species trees. **Systematic biology**, v. 46, n. 3, p. 523-536, 1997.
- MAJURE, L. C. et al. Phylogeny of *Opuntia* ss (Cactaceae): clade delineation, geographic origins, and reticulate evolution. **American Journal of Botany**, v. 99, n. 5, p. 847-864, 2012.
- MCCORMACK, John E. et al. Applications of next-generation sequencing to phylogeography and phylogenetics. **Molecular Phylogenetics and Evolution**, v. 66, n. 2, p. 526-538, 2013.
- MENEZES, M. O. T. et al. Pleistocene radiation of coastal species of *Pilosocereus* (Cactaceae) in eastern Brazil. **Journal of Arid Environments**, v. 135, p. 22-32, Dez. 2016.
- MINH, B. Q. et al. Ultrafast Approximation for Phylogenetic Bootstrap. **Molecular Biology Evolution**, v. 30, n. 5, p. 1188-1195, Mai. 2013.
- MORAES, Evandro M. et al. Population genetic structure of two columnar cacti with a patchy distribution in eastern Brazil. **Genetica**, v. 125, n. 2-3, p. 311-323, 2005.
- NEGRÓN-ORTIZ, V. Chromosome numbers, nuclear DNA content, and polyploidy in *Consolea* (Cactaceae), an endemic cactus of the Caribbean Islands. **American Journal of Botany**, v. 94, n. 8, p. 1360–1370, Ago. 2007.
- NGUYEN, K. T. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. **Molecular Biology Evolution**, v. 32, n. 1, p. 268-74, Jan. 2015.
- NIELSEN, R. BEAUMONT, M. A. Statistical inferences in phylogeography. **Molecular Ecology**, v. 18, n. 6, p. 1034–1047, Mar. 2009.
- NYFFELER, R. Phylogenetic relationships in the cactus family (Cactaceae) based on evidence from trnK/matK and trnL-trnF sequences. **American Journal of Botany**, v. 89, n. 2, p. 312-326, Fev. 2002.
- PEREZ, M. F. et al. Model-based analysis supports interglacial refugia over long-dispersal events in the diversification of two South American cactus species. **Heredity**, v. 116, n. 6, p. 550-557, Jun. 2016a.

PEREZ, M. F. et al. Anonymous nuclear markers reveal taxonomic incongruence and long-term disjunction in a cactus species complex with continental-island distribution in South America. **Molecular Phylogenetics and Evolution**, v. 95, p. 11-19, Feb. 2016b.

PETTENGILL, J. B. et al. An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: a case study with *Salmonella*. **PeerJ**, DOI 10.7717/peerj.620, Out. 2014.

PETERSON, B. K. et al. Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-Model Species. **Plos One**, v. 7, n. 5, Mai. 2012.

PETIT, R. J; EXCOFFIER, L. Gene flow and species delimitation. **Trends in Ecology and Evolution**, v. 24, n. 7, p. 386–393, Jul. 2009.

POSADA, D. jModelTest: phylogenetic model averaging. **Molecular Biology Evolution**, v. 25, n. 7, Jul. 2008.

PURITZ, J. B. et al. Demystifying the RAD fad. **Molecular Ecology**, v. 23, p. 5937–5942, Jun. 2014.

REDDY, S. et al. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. **Systematic Biology**, v. 66, n. 4, p. 857–879, Mar. 2017.

RHOADS, K; AU, F. PacBio Sequencing and Its Applications. **Genomics, Proteomics & Bioinformatics**, v. 13, n. 5, p. 278-289, Out. 2015.

RITZ, C. M. et al. The molecular phylogeny of *Rebutia* (Cactaceae) and its allies demonstrates the influence of paleogeography on the evolution of south american mountain cacti. **American Journal of Botany**, v. 94, n. 8, p. 1321–1332, 2007.

ROMEIRO-BRITO, M. et al. Lineage-specific evolutionary rate in plants: Contributions of a screening for *Cereus* (Cactaceae). **Applications in plant sciences**, v. 4, n. 1, p. 1-8, Jan. 2016.

ROMEIRO-BRITO, M. **Filogenia multiloco do gênero *Pilosocereus*** (Cactaceae, Cereae). Sorocaba: Universidade federal de São Carlos, 2017, 93p.

RUBIN, B. E. R; REE, R. H; MOREAU, S. C. Inferring Phylogenies from RAD Sequence Data. **Plos One**, v. 7, n. 4, Abr. 2012.

SCHAAL, A. B; OLSEN, K. M. Gene genealogies and population variation in plants. **PNAS**, v. 97, n. 13, p. 7024-7029, Jun. 2000.

SHAW, J. et al. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. **American journal of botany**, v. 92, n. 1, p. 142-166, Jan. 2005.

SHAW, J. et al. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. **American journal of botany**, v. 101, n. 11, p. 1987-2004, 2014.

SMALL, R. L; CRONN, R. C; WENDEL, J. F. Use of nuclear genes for phylogeny reconstruction in plants. **Australian Systematic Botany**, v. 17, n. 2, p. 145-170, 2004.

SOUSA, F et al. Phylogenetic Properties of 50 Nuclear Loci in Medicago (Leguminosae) Generated Using Multiplexed Sequence Capture and Next-Generation Sequencing. **Plos One**, v. 10, n. 5, Mai. 2014.

STAMATAKIS, A; HOOVER , P; ROUGEMONT, J. A Rapid Bootstrap Algorithm for the RAxML Web Servers. **Systematic Biology**, v. 57, n. 5, p. 758–771, Out. 2008.

STAMATAKIS, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics**, v. 30, n. 9, Mai. 2014.

STRAUB, S. C. K. et al. Navigating the tip of the genomic iceberg: Next-Generation Sequencing for Plant Systematics. **American Journal of Botany**, v. 99, n. 2, p. 349-364, Fev. 2012.

STREICHER, J, W; SCHULTE, J. A; WIENS, J. J. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in Iguanian Lizards. **Systematic Biology**, v. 65, n. 1, p. 128–145, Ago. 2016.

TAYLOR, N.P; ZAPPI, D. *Cacti of Eastern Brazil*. Royal Botany Garden, Reino Unido, 2004.

TURCHETTO-ZOLET, A. C. et al. Phylogeographical patterns shed light on evolutionary process in South America. **Molecular Ecology**, v. 22, n. 10, p. 1193-1213, Mar. 2013.

VÁZQUEZ-SÁNCHEZ, M. et al. Molecular phylogeny, origin and taxonomic implications of the tribe Cactaeae (Cactaceae). **Systematics and Biodiversity**, v. 11, p. 103–116, Abr. 2013.

XIA, Z. et al. AFSM sequencing approach: a simple and rapid method for genome-wide SNP and methylation site discovery and genetic mapping. **Scientific Reports**, DOI: 10.1038, Dez. 2014.

WAGNER, C. E. et al. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. **Molecular Ecology**, v. 22, p. 787–798, 2013.

WELCH, A. J. et al. The quest to resolve recent radiations: Plastid phylogenomics of extinct and endangered Hawaiian endemic mints (Lamiaceae). **Molecular phylogenetics and evolution**, v. 99, p. 16-33, Mar. 2016.

YANG, G. Q. et al. Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. **Plant Methods**, DOI 10.1186/s13007-016-0139-1, Ago. 2016.

ZAPPI, D.; TAYLOR, N.; LAROCCA, J. Riqueza das Cactaceae no Brasil. In: *Plano de Ação Nacional para a Conservação das Cactáceas*. Org. Silva, S.R. - Instituto Chico Mendes de Conservação da Biodiversidade, ICMbio, 2011.

ZAPPI, D.; TAYLOR, N.; SANTOS, M. r. Riqueza das Cactaceae no Brasil. In: *Plano de Ação Nacional para a Conservação das Cactáceas*. Org. Silva, S.R. - Instituto Chico Mendes de Conservação da Biodiversidade, ICMbio, 2011.

ZELLMER, A. J. et al. Deep Phylogeographic Structure and Environmental Differentiation in the Carnivorous Plant **Sarracenia alata**. **Sustematic Bioligy**, v. 61, n. 5, p. 763–777, Abr. 2012.

ZHBANNIKOV, I. Y. et al. SeqyClean: A Pipeline for High-throughput Sequence Data Preprocessing. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB '17). ACM, New York, NY, USA, 407-416. DOI: <https://doi.org/10.1145/3107411.3107446>, 2017.

ZHOU, X. et al. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. **Molecular Biology and Evolution**, doi:10.1093/molbev/msx302, Nov. 2017.

ZIMMER, A. E; WEN, J. Using nuclear gene data for plant phylogenetics: Progress and prospects. **Molecular Phylogenetics and Evolution**, v. 65, p. 774-785, Jul. 2012.

APÊNDICES

APÊNDICE A: Protocolo ddRAD-Seq (Biblioteca I)

Foi preparada uma biblioteca reduzida com 50 indivíduos da Tabela 1. Para tanto foi adaptado o protocolo ddRAD-Seq descrito em Kess et al. (2016). Para este protocolo utilizamos as enzimas *EcoRI* (NEB) e *MseI* (NEB). Foi utilizado um total de 90ng do DNA genômico de cada indivíduo, que foi então digerido simultaneamente com 0.02 U de cada uma das enzimas [*MseI* (10U/ μ L) e *EcoRI* (20U/ μ L)]. As digestões ocorreram por 2h a 37°C, seguida da inativação de 20 minutos a 65°C.

Para a reação de ligação, adaptadores fita-dupla (APÊNDICE B), foram ligados ao produto de digestão por meio de uma reação com 1 μ M do adaptador *MseI*, 0,05 μ M do adaptador da *EcoRI*, 6 μ L do produto de digestão, Tampão T4 *Ligase* Buffer 1X, e 40U da enzima T4 DNA *Ligase*. Essa reação foi incubada por 2h a 22°C, seguida pela inativação de 20 minutos a 65°C.

Os identificadores (indexes) i5 e i7 padronizados pela Illumina (APÊNDICE B) e a sequência complementar da “*flow-cell*” Illumina foram inseridos nos nos insertos ligados aos adaptadores por meio de uma PCR de fusão com primers contendo sequencias complementares aos adaptadores *MseI* e *EcoRI*, além dos indexes i5 e i7 e a sequência da *flow-cell* Illumina. Para tanto, foi montada uma reação contendo 3 μ L do produto de ligação, 0.5 μ M de cada primer de fusão, 200 μ M de DNTP, 0.6 μ L de DMSO e 2U da enzima de alta-fidelidade Phusion® DNA *Polimerase* juntamente com seu Tampão (NEB, Inc.). Os *primers* foram combinados em cada reação de forma que cada amostra tivesse uma combinação única. A ciclagem da reação de PCR envolveu 17 ciclos com um passo inicial de 98°C por 30 segundos, um passo de desnaturação (98°C por 20 segundos), um passo de anelamento (55°C por 30 segundos), um passo de extensão (72°C por 40 segundos) e um passo de extensão final a 72°C por 10 minutos. A reação de PCR foi realizada em triplicata para cada amostra visando aumentar a concentração de DNA final e minimizar a amplificação preferencial de alelos (PETERSON et al. 2012).

O produto da PCR (2 uL) foi visualizado em gel de agarose 1 %, com uma corrida de 2h a 70 V, com o uso do marcador de peso molecular *Low Mass DNA Ladder* (Invitrogen) e corado com *Invitrogen™ SYBR™ Safe DNA Gel Stain*. As reações de PCR de por amostra

combinadas foram purificadas com o conjunto de reagentes *DNA Clean & Concentrator Zymo* e eluídas em um volume final de 12µL, dos quais 2µL foram usados para quantificação no *Qubit 3.0 Fluorometer* (Invitrogen) com o *QubitTM dsDNA HS Assay Kit*. A partir dessa quantificação as amostras foram normalizadas e uma solução de *pool*, com as amostras em condições equimolares, foi produzida.

Após as simulações no SimRAD, foi selecionado um intervalo de seleção de tamanho entre 300pb e 400pb, gerando uma cobertura de aproximadamente 13x, ou seja, 13 leituras por loco. O gel utilizado para seleção de tamanho foi de 1,5% de agarose e foi corado com o intercalante *InvitrogenTM SYBRTM Safe DNA Gel Stain* após o fim da corrida eletroforética, buscando minimizar os efeitos que intercalantes podem causar na migração do DNA. O fragmento selecionado do gel, foi purificado (*QIAquick Gel Extraction Kit*), eluído em um volume final de 240uL e posteriormente concentrado para 50uL.

O *pool* gerado após a PCR foi analisado por meio do *Bioanalyzer* e de PCR em tempo real (qPCR). O sequenciamento foi realizado em um sequenciador *HiSeq 2500 (High Output Run Mode)*, onde a preparação do *pool* foi realizada com o *Illumina TruSeq Strandedm RNA Sample Prep LS Protocol*, e o sequenciamento foi feito em uma *HiSeq Flow Cell v4*, com o *HiSeq SBS Kit v4* e leituras pareadas de 150 bp (2x).

As análises de controle de qualidade dos dados das foi realizada nos programas FastQC (disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>). Para a remoção das sequencias de adaptadores e outros contaminantes que poderiam estar presentes na biblioteca, foi utilizado o software SeqyClean 1.9.10 (ZHBANNIKOV et al. 2017) em duas situações distintas: (1) considerando os dados de maneira *paired-end* (PE) e (2) considerando os dados de maneira *single-end* (SE).

O pacote PyRAD (EATON, 2014) foi usado para a chamada de SNPs após a filtragem no SeqyClean e os resultados podem ser observados na Tabela 1.

Tabela 1- Números de reads gerados pelo sequenciamento HiSeq e após as filtrações (SeqClean e PyRAD), além do número de loci gerados em cada uma das estratégias de triagens (PE e SE).

Amostras	Análise PE			Análise SE		
	Nº de leituras analisadas	Nº de leituras após filtração	Nº Loci	Nº de leituras analisadas	Nº de leituras após filtração	Nº Loci
S149V11	2132028	1106616	145	669414	588536	313
S149V15	7054946	3306308	174	2492082	2128549	412
S44V3	6854694	3383650	162	2339798	1967766	418
S38V7B	5760498	2989900	175	2167364	1921244	427
S113F9	9489050	4594214	176	3337810	2924975	434
S113F7	4863602	2424082	165	1470922	1284370	431
S110F4	4283336	2023110	175	1372104	1127259	417
S90V3	5283818	2319164	174	1585176	1349213	422
S103C1	7655958	3457228	176	2334282	1995644	423
S103E1	6319368	2959436	173	1966208	1677578	428
S52V3	1722516	781538	135	457160	386744	196
S144F2	5746130	2232624	168	1477272	1224473	402
S77A7	2770912	1311272	145	807052	710798	315
S145F4	8280058	2816918	179	2060350	1749618	435
S72F11	7280882	3341664	169	2414132	2025548	435
S72F12	6607628	3261936	175	2334926	2065380	440
S80F8	5590964	2817408	172	1835022	1576491	431
S114F16	5080324	2433484	181	1650142	1373203	434
S114F8	4948718	2440516	176	1783288	1498190	437
S104F4	7089114	2922048	183	2002638	1701219	432
S104F17	5828158	2778582	180	1950704	1662499	432
S76F10	5464390	2006938	168	1390572	1132857	374
S76F6	2240956	988922	145	601456	504383	247
S85F3	2870466	1376026	140	807342	701672	317
S85F2	8657536	4184074	181	3199446	2694420	434

Tabela 1. Conclusão

Amostras	Análise PE			Análise SE		
	Nº de leituras analisadas	Nº de leituras após filtragem	Nº Loci	Nº de leituras analisadas	Nº de leituras após filtragem	Nº Loci
S88F3	6787948	3405122	168	2294794	1911856	430
S88F4	5858778	3020582	172	2088340	1832953	428
S77A24	8418324	4299918	174	3012964	2608074	435
S115C41	5115774	2642736	162	1642586	1421054	429
S115B21	4163490	1953804	171	1341060	1093783	414
S115A1	5409802	2475616	174	1729996	1460982	435
S149V14	7855842	3224838	179	2082960	1755801	424
S77A31	4558034	2042860	173	1330440	1127586	397
S77A30	1845494	747584	114	439076	367415	174
S103D1	6316580	1882080	143	1109332	864154	324
S99A6	2081466	893632	156	470354	408768	348
S131F4	4945720	2153730	170	1417138	1194016	413
S150F1	5497690	2907556	83	1940762	1606259	239
S150F2	4685486	2352050	144	1609160	1412257	386
S149V16	7368954	3873620	144	2801402	2423057	384
S10B6	4906950	2116752	155	1265770	1097292	405
S137A1	3658250	1616518	159	1045886	847804	386
S77A18	4155074	1971828	166	1313752	1105178	410
S147F6	5343528	2098768	159	1410814	1152786	390
S77A17	4416706	2135842	178	1428872	1218527	421
S38V6B	1544048	622712	118	382878	321620	217
S26A16	4156780	1483942	134	915272	727500	300
S62A19	4017010	1707814	138	1075060	910552	347
S35M1	1885274	914014	113	507084	445005	251

O número de sequências analisadas diz respeito as sequências após a filtragem no SeqyClean.

APÊNDICE B - Sequências de adaptadores e *primers* utilizados para a preparação das bibliotecas (I e II).

Biblioteca	Identificação	Sequência
I	Adaptador P1 (<i>EcoRI</i>)	a 5' ACACCTCTTTCCCTACACGACGCTCTTCCGATCT 3'
		b 3' AATTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT 5'
	Adaptador P2 (<i>MseI</i>)	a 5' TAAGATCGGAAGAGCGAGAACAA 3'
		b 3' GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT 5'
	PCR primers Nextera® XT Index Kit	P5 5' AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGAC 3'
		P7 5' CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGC 3'
II	Adaptador P1 (<i>EcoRI</i>)	a 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG 3'
		b 3' AGCAGCCGTCGCAGTCTACACATATTCTCTGTCTTAA 5'
	Adaptador P2 (<i>HPAII</i>)	a 5' CGCTGTCTCTTATACGTCCTCTT 3'
		b 3' GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG5'
	PCR primers Nextera® XT Index Kit	P5 5' AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC 3'
		P7 5' CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGC 3'

APÊNDICE C - Arquivo de parâmetros utilizado para controlar as opções durante o *SNP calling* no programa iPyRAD 0.7.20. Em destaques (círculos em vermelho) a opções mais relevantes que foram modificadas pelo utilizador. A abreviação DF% na opção 21 se refere ao controle de porcentagem de dados faltantes permitidos em cada um dos cenários: até 40% (mínimo de 39 amostras), até 60% (mínimo de 26 amostras) e até 80% (mínimo de 13 amostras).

```

----- ipyrad params file (v.0.7.20)-----
name                ## [0] [assembly_name]: Assembly name. Used to name output directories for assembly steps
./                  ## [1] [project_dir]: Project dir (made in curdir if not present)
                    ## [2] [raw_fastq_path]: Location of raw non-demultiplexed fastq files
                    ## [3] [barcodes_path]: Location of barcodes file
input/*.fastq.gz    ## [4] [sorted_fastq_path]: Location of demultiplexed/sorted fastq files
denovo              ● ## [5] [assembly_method]: Assembly method (denovo, reference, denovo+reference, denovo-reference)
                    ## [6] [reference_sequence]: Location of reference sequence file
pairedddrad         ● ## [7] [datatype]: Datatype (see docs): rad, gbs, ddrad, etc.
AATTC, CGG         ● ## [8] [restriction_overhang]: Restriction overhang (cut1,) or (cut1, cut2)
5                  ## [9] [max_low_qual_bases]: Max low quality base calls (Q<20) in a read
33                 ● ## [10] [phred_Qscore_offset]: phred Q score offset (33 is default and very standard)
5                 ● ## [11] [mindepth_statistical]: Min depth for statistical base calling
6                 ● ## [12] [mindepth_majrule]: Min depth for majority-rule base calling
10000              ● ## [13] [maxdepth]: Max cluster depth within samples
0.85               ● ## [14] [clust_threshold]: Clustering threshold for de novo assembly
0                 ● ## [15] [max_barcode_mismatch]: Max number of allowable mismatches in barcodes
2                 ● ## [16] [filter_adapters]: Filter for adapters/primers (1 or 2=stricter)
65                ● ## [17] [filter_min_trim_len]: Min length of reads after adapter trim
2                 ● ## [18] [max_alleles_consens]: Max alleles per site in consensus sequences
5, 5              ● ## [19] [max_Ns_consens]: Max N's (uncalled bases) in consensus (R1, R2)
6, 6              ● ## [20] [max_Hs_consens]: Max Hs (heterozygotes) in consensus (R1, R2)
DF%               ● ## [21] [min_samples_locus]: Min # samples per locus for output
20, 20            ● ## [22] [max_SNPs_locus]: Max # SNPs per locus (R1, R2)
8, 8              ● ## [23] [max_Indels_locus]: Max # of indels per locus (R1, R2)
0.5               ● ## [24] [max_shared_Hs_locus]: Max # heterozygous sites per locus (R1, R2)
0, 0, 0, 0        ● ## [25] [trim_reads]: Trim raw read edges (R1>, <R1, R2>, <R2) (see docs)
0, 0, 0, 0        ● ## [26] [trim_loci]: Trim locus edges (see docs) (R1>, <R1, R2>, <R2)
*                 ● ## [27] [output_formats]: Output formats (see docs)
                    ● ## [28] [pop_assign_file]: Path to population assignment file

```

APÊNDICE D – Árvore filogenética Máxima Verossimilhança baseada na matriz SNPs ligados contendo 3694 sítios e permitindo até 40% de dados faltantes gerada no programa IQTree. Os suportes de ramos foram considerados significativos quando maiores que 80 para Shimodaira-Hasegawa (SH-aLRT) e maiores que 95 para o *ultrafast bootstrap* (UFBOOT), representados nessa ordem na árvore. Os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura.

APÊNDICE F – Árvore filogenética Máxima Verossimilhança baseada na matriz Loci contendo 333 loci e permitindo até 40% de dados faltantes gerada no programa IQTree. Os suportes de ramos foram considerados significativos quando maiores que 80 para Shimodaira-Hasegawa (SH-aLRT) e maiores que 95 para o *ultrafast bootstrap* (UFBOOT), representados nessa ordem na árvore. Os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura.

APÊNDICE G – Árvore filogenética de Máxima Verossimilhança baseada na matriz Loci contendo 333 loci e permitindo até 40% de dados faltantes gerada no programa RaXML. Os suportes de ramos foram considerados significativos quando maiores que 70 para o *rapid bootstrap* (RBS). Os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura.

APÊNDICE I – Árvore filogenética de Máxima Verossimilhança baseada na matriz SNPs ligados contendo 21.442 sítios e permitindo até 60% de dados faltantes gerada no programa RaXML. Os suportes de ramos foram considerados significativos quando maiores que 70 para o *rapid bootstrap* (RBS). Os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura.

APÊNDICE J – Árvore filogenética Máxima Verossimilhança baseada na matriz Loci contendo 1440 loci e permitindo até 60% de dados faltantes gerada no programa IQTree. Os suportes de ramos foram considerados significativos quando maiores que 80 para Shimodaira-Hasegawa (SH-aLRT) e maiores que 95 para o *ultrafast bootstrap* (UFBOOT), representados nessa ordem na árvore. Os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura.

APÊNDICE K – Árvore filogenética de Máxima Verossimilhança baseada na matriz Loci contendo 1440 loci e permitindo até 60% de dados faltantes gerada no programa RaXML. Os suportes de ramos foram considerados significativos quando maiores que 70 para o *rapid bootstrap* (RBS). Os suportes de ramos não significativos são representados pelo sinal negativo (-). O sinal de interrogação (?) indica dúvida sobre a identificação da espécie na literatura.