UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BOTÂNICA
CURSO NOTURNO DE LICENCIATURA EM CIÊNCIAS BIOLÓGICAS

Victor Soares Santibañez

Avaliação de marcadores de sequenciamento RAD para inferência filogenética de máxima verossimilhança no gênero *Barbacenia* (Velloziaceae)

Florianópolis
2019

Victor Soares Santibañez

Avaliação de marcadores de sequenciamento RAD para inferência filogenética de máxima verossimilhança no gênero *Barbacenia* (Velloziaceae)

Trabalho Conclusão do Curso de Graduação em Licenciatura em Ciências Biológicas do Centro de Ciências Biológicas da Universidade Federal de Santa Catarina como requisito para a obtenção do Título de Licenciado em Ciências Biológicas
Orientador: Profa. Dra. Suzana de Fátima Alcantara

Florianópolis
2019

Ficha de identificação da obra

Santibanez, Victor
Avaliação de marcadores de sequenciamento RAD para inferência filogenética de máxima verossimilhança no gênero Barbacenia (Velloziaceae) / Victor Santibanez ; orientadora, Suzana Alcantara, 2020.
59 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Graduação em Ciências Biológicas, Florianópolis, 2020.

Inclui referências.
1. Ciências Biológicas. 2. sistemática filogenética. 3. bioinformática genômica. 4. botânica. I. Alcantara, Suzana. II. Universidade Federal de Santa Catarina. Graduação em Ciências Biológicas. III. Título.

Victor Soares Santibañez

Avaliação de marcadores de sequenciamento RAD para inferência filogenética de máxima verossimilhança no gênero *Barbacenia* (Velloziaceae)

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de Licenciado em Ciências Biológicas e aprovado em sua forma final pelo Curso de Licenciatura em Ciências Biológicas.

Florianópolis, 11 de dezembro de 2019.

_____

Prof. Carlos Roberto Zanetti, Dr.

Coordenador do Curso

Banca Examinadora:

_____

Prof.ª Dra. Suzana de Fátima Alcantara

Orientadora

Universidade Federal de Santa Catarina

_____

Dra. Duane Fernandes Lima

Universidade Federal de Santa Catarina

_____

Prof. Dr. Guilherme de Toledo e Silva

Universidade Federal de Santa Catarina

Dedico este trabalho aos meus pais Marcos Rinaldo Santibañez Arias e Carme Luci Soares Santibañez.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, irmão e sobrinha pela apoio infinito e por perdoar a minha ausência durante estes anos de curso. Agradeço aos meus amigos pelo ombro oferecido.

Agradeço aos meus colegas do grupo de discussão PLENTBio do Departamento de Botânica pela companhia na exploração da sistemática vegetal, e principalmente a minha orientadora Suzana Alcantara por me ensinar mais sobre sistemática, bioinformática e fazer ciência do que todo o resto do curso combinado. Agradeço também pela sua acessibilidade, respeito e empenho em me ajudar a concluir este trabalho.

Agradeço aos pesquisadores Isabel Gerhardt e Ricardo Dante do departamento de informática da Embrapa pela colaboração e disponibilização do genoma de *Vellozia nivea*, além do pesquisador Renato de Mello-Silva da Universidade de São Paulo por algumas das sequências Sanger utilizadas neste trabalho.

Agradeço à Universidade Federal de Santa Catarina por possibilitar este trabalho e também ao Instituto Serrapilheira pelo apoio financeiro dado a esta pesquisa. (GRANT SERRA-1709-21213).

“Exatamente!" disse o Pensador Profundo. “Então, uma vez que você realmente souber qual é a pergunta, você saberá o que a resposta significa.” (Douglas Adams, O Guia do Mochileiro das galáxias, 1979)

RESUMO

Neste trabalho, testei marcadores obtidos a partir do sequenciamento RAD (*Restricted site-associated DNA sequencing*) para a inferência de filogenias em um clado de plantas tropicais. Todas as análises foram feitas com a utilização de *pipelines* de montagem de genomas e de inferência filogenética de acesso livre. Para isso, amostras de sequenciamento RAD de 31 espécies do gênero *Barbacenia*, previamente coletadas e sequenciadas, foram analisadas. Os alinhamentos foram testados com parâmetros distintos e os diferentes *datasets* resultantes foram analisados para a reconstrução de uma árvore filogenética com método de inferência de máxima verosimilhança. As árvores filogenéticas resultantes de cada alinhamento foram comparadas para verificar os parâmetros como sustentação e comprimento dos ramos. As árvores inferidas somente com base nos marcadores RAD foram então comparadas com o conhecimento científico mais atual sobre a família, que baseia-se em quatro sequências inter-gênicas obtidas pelo método de sequenciamento Sanger. Com isso, foi possível identificar resultados discrepantes entre os marcadores, a partir dos quais o potencial de utilização de marcadores RAD foi avaliado, especialmente no que diz respeito à resolução das relações referentes aos nós mais profundos da filogenia em questão.

Palavras-chave: Bioinformática. Marcadores genéticos. Sistemática molecular.

ABSTRACT

I tested the precision of markers obtained by RAD (Restricted site-associated DNA) sequencing to infer phylogenies in a clade of tropical plants by the utilization of freely accessible pipelines. For that, samples of 31 species of the genus *Barbacenia*, previously collected and sequenced, were analyzed using assemblage pipelines. This process was tested with different parameters and the resulting datasets were analyzed to infer a phylogenetic tree by the method of maximum likelihood. The resulting trees were compared to verify parameters such as branch length and node support. The trees inferred using RAD markers only were then compared to the current knowledge of the group, which is based on 4 intergenic sequences obtained by the Sanger sequencing method. By evaluating the obtained trees and the combined tree it was possible to identify possible discrepancies within the methods. These analyses showed the potential of using RAD markers to elucidate phylogenetic relationships, especially regarding the most ancient nodes within this group.

Keywords: Bioinformatics. Genetic markers. Molecular systematics.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RAD	*Restricted site-associated DNA sequencing*
snps	*single nucleotide polimorfims*

SUMÁRIO

1 INTRODUÇÃO

Com o estabelecimento da teoria evolutiva a partir da publicação da obra de Darwin em 1859, passamos a entender que todos os seres vivos fazem parte de uma grande árvore evolutiva. Desvendar a configuração dos galhos desta árvore, ou o grau de parentesco entre os seres vivos, passou então a ser um dos desafios da ciência. A importância de criar filogenias precisas e completas é visível em várias áreas das ciências biológicas, dado que identificar organismos é necessário para esforços de conservação, produção de insumos de origem biológica identificação de patógenos. Para a classificação e organização dos seres vivos, nasceu uma área da ciência chamada sistemática filogenética, que tem entre seus objetivos condensar e resumir a história da evolução dos organismos vivos que representa a melhor hipótese concebível com os dados disponíveis, e que tenta recuperar partes da árvore da vida a partir do estudo das similaridades e diferenças entre organismos (MORRISON, 2012). Para a pesquisa bioprospectiva, a busca por novos genes ou metabólitos é guiada pelo parentesco entre as espécies, já que espécies mais próximas tendem a produzir metabólitos semelhantes e compartilhar genes (KERSTEN; WENG, 2018).

Durante muitas décadas, a tarefa de montar árvores filogenéticas foi limitada pela tecnologia disponível, uma vez que isso dependia de comparar homologias morfológicas existentes nos organismos vivos para inferir suas histórias evolutivas (UNDA, 2005). Com o advento do sequenciamento genético, os sistematas e biólogos evolutivos passaram a ter ao seu alcance uma nova ferramenta para inferir as relações de parentesco entre organismos representando diferentes espécies. A utilização de computadores para manipular e investigar dados biológicos, especialmente de biologia molecular, deu origem a área de estudo chamada bioinformática (LUSCOMBE; GREENBAUM; GERSTEIN, 2001). Esta ciência permite a análise de informação genética e proteica sobre um olhar evolutivo, considerando estudos de outras áreas da biologia como a paleontologia e arqueologia, para inferir o parentesco entre linhagens. Enquanto a sistemática clássica depende da comparação de homologias morfológicas macro e microscópicas, os métodos moleculares fazem algo semelhante, mas com o material genético. Após o sequenciamento, o material obtido é comparado entre diferentes táxons na busca de regiões ortólogas, e então, as diferenças entre estas regiões guiam a inferência estatística da filogenia mais provável (MORRISON, 2012). Para que isso seja possível, um grande conhecimento teve que ser desenvolvido não só na criação de

dispositivos de sequenciamento precisos e práticos, mas também na manipulação destes dados durante todos os passos do processo de sequenciamento e inferência filogenética.

Para inferir a relação de parentesco entre os táxons sequenciados, uma exaustiva comparação das sequências de DNA é feita para inferir uma árvore através de uma análise probabilística. Em um mundo ideal com infinita capacidade de processamento, avaliaríamos todas as combinações possíveis destes dados, na tentativa de encontrar a combinação que melhor se encaixa ao modelo utilizado. No mundo real por outro lado a limitação computacional é um problema que só aumenta com a quantidade de dados a serem considerados (MORRISON, 2012). Para resolver este problema, foram desenvolvidos métodos que permitem estimar a história evolutiva mais provável com um uso limitado de recursos computacionais. Uma destas análises é a análise de máxima verosimilhança, que usa uma abordagem baseada em critérios (BALDING; MOLTKE; MARIONI, 2019). Neste tipo de análise, o modelo que melhor simula a probabilidade de alteração nas bases nitrogenadas é utilizado para guiar a busca. Basicamente, o modelo possibilita o cálculo da probabilidade de transição Pij(t), que é a probabilidade que um estado j vai existir ao final de um ramo de comprimento t, se o estado inicial for i. Este método toma duas suposições. A primeira é que a evolução de sítios diferentes em uma dada árvore é independente, e a segunda é que a evolução em diferentes linhagens é independente. Isso significa que é possível calcular a probabilidade para cada sítio e comparar as probabilidades posteriormente (FELSENSTEIN, 2003). Para economizar poder computacional, o método de amostragem de subárvore é utilizado. Ele utiliza uma variável chamada verossimilhança condicional de uma subárvore. O processo de busca começa nas extremidades da árvore, onde a probabilidade é mais fácil de computar. A análise se move árvore abaixo, calculando a probabilidade de cada nó dadas as probabilidades anteriores. Ao chegar na raiz da árvore, o algoritmo terá computado a probabilidade de todas as observações. A amostragem de subárvore permite economizar poder computacional ao utilizar valores de probabilidade de ramo já calculados em diferentes nós, eliminando a necessidade de calcular todo o ramo todas as vezes (BALDING; MOLTKE; MARIONI, 2019). Este processo é repetido diversas vezes e possíveis combinações são simuladas para encontrar a árvore com a topologia com as maiores chances de acomodar os dados com base no modelo. Para isso, um valor de verossimilhança é gerado e comparado para cada árvore, considerando o número de alterações em comum necessárias para tal

agrupamento. Está variável é chamada de comprimento de ramo, e é importante pois quanto maior, menor a proximidade entre os terminais (MORRISON, 2012).

Além da limitação dos algoritmos, segundo Rubin, Ree e Moreau (2012), a sistemática molecular esteve limitada por muitos anos a utilizar algumas poucas sequências de genes ortólogos para inferir relações filogenéticas. Esta limitação se deu pela dificuldade em encontrar estes genes que apresentassem variabilidade suficiente para resolver relações filogenéticas com confiança e pudessem ser sequenciados e amplificados para este fim. Uma solução para este problema é o sequenciamento de leituras curtas do tipo RadSeq (*Restricted site-associated DNA sequencing*, chamado apenas de RAD a partir daqui), que tem como alvo os sítios comuns de restrição enzimática que são encontrados por todo genoma, permitindo uma amostragem maior de marcadores associados a estes locais de restrição (EATON; REE, 2013). Por ser capaz de encontrar muitos polimorfismos de nucleotídeo único, este tipo de sequenciamento foi utilizado inicialmente em estudos para encontrar variações em nível de população, e gradualmente esta ferramenta está sendo voltada para estudos filogenéticos (EATON; REE, 2013). Um fator que gera incerteza quanto à aplicabilidade do sequenciamento do tipo RAD para resolver filogenias são os dados faltantes, que podem ser encontrados em matrizes de sequenciamento do tipo RAD. Apesar disso, Eaton et al. (2016) apontam com simulações que a consideração cuidadosa na seleção dos métodos de alinhamento e do cálculo de inferência permitem a inferência de filogenias precisas com alto nível de sustentação estatística. Podemos inclusive encontrar estudos utilizando esta ferramenta e semelhantes para inferir filogenias de diversas áreas da biologia (HOU et al., 2015; SATLER et al., 2019).

Nos trópicos concentra-se um grande número de espécies vegetais e a investigação de filogenias de plantas tropicais pode ser grandemente favorecida pela utilização de marcadores do tipo RAD. A grande diversidade dos trópicos, influenciada tanto pela ocorrência de radiação evolutiva quanto por fatores ecológicos atuais, faz com que o estudo de linhagens tropicais ainda seja dificultado pelos custos de obtenção de marcadores (CHOI et al., 2019). Além disso, processos evolutivos em curso, como especiação incipiente e/ou radiação recente, também dificultam a obtenção de marcadores informativos, acarretando custos e dificuldades no desenvolvimento desses marcadores. Nesse contexto, a utilização de marcadores do tipo RAD pode ser particularmente informativa para a resolução de filogenias problemáticas. Um exemplo de filogenia pouco resolvida são aquelas que vêm sendo obtidas

para o gênero *Barbacenia* (família Velloziaceae, monocotiledônea da ordem Pandanales). A família Velloziaceae tem se mostrado especialmente difícil de resolver, com a alteração e dissolução de grupos que se mostraram parafiléticos após análise mais profunda de suas supostas homologias (MELLO-SILVA et al., 2011). O grupo *Barbacenia* é o maior exemplo disso, tendo recentemente incluído os gêneros parafiléticos *Aylthonia*, *Burlemarxia* e *Pleurostima*, que se mostraram ter sido agrupados por homoplasias (MELLO-SILVA et al., 2011). Este grupo endêmico da América do Sul possui em torno de 100 espécies, encontradas na Amazônia, Caatinga, Cerrado e Mata Atlântica (MELLO-SILVA et al., 2011), estando presentes das vegetações de campo de altitude, campo de várzea, campo rupestre e vegetação sobre afloramentos rochosos (MELLO-SILVA et al., 2011). O gênero *Barbacenia* possui sua distribuição quase que inteiramente em regiões rochosas, e possui membros que apresentam tolerância à dessecação (GAFF, 1977) e resistência à seca (AIDAR et al., 2010; ALCANTARA et al., 2015). Atualmente a circunscrição deste grupo está melhor estabelecida (MELLO-SILVA et al., 2011). Com análises feitas por Alcantara, Ree e Mello-Silva (2018) contendo um maior número de terminais, foi possível verificar que o gênero mantém alta sustentação como monofilético, apesar das relações infra-genéricas ainda possuírem pouca sustentação no geral, principalmente nos nós mais profundos dentro de *Barbacenia*.

Nesse contexto, este trabalho de conclusão de curso é um estudo científico em que amostras obtidas por sequenciamento RAD foram processadas a partir de *scripts* de bioinformática e utilizadas para a inferência de uma árvore filogenética do gênero *Barbacenia*. A árvore resultante da análise dos marcadores RAD foi comparada com a filogenia molecular mais recente disponível, para testar a compatibilidade e a possibilidade de combinação entre os marcadores em uma análise de evidência total molecular. Isso possibilitou avaliar a eficácia dos marcadores obtidos a partir do sequenciamento RAD no cenário da botânica de plantas tropicais, na qual diferentemente de trabalhos semelhantes desenvolvidos com famílias de ambientes temperados (FITZ-GIBBON et al., 2017), existem desafios como uma maior riqueza de espécies e altos índices de endemismo.

1.1 JUSTIFICATIVA

Considerando o histórico problemático da classificação taxonômica do grupo, reconstruir uma filogenia bem resolvida será útil para subsidiar avanços em estudos

sistemáticos e evolutivos de Velloziaceae e aumentar o conhecimento das relações evolutivas em *Barbacenia*. Além disso, por apresentar resistência a dessecação, é de interesse científico e comercial a busca por genes que possam oferecer esta mesma resistência para futura manipulação genética com culturas de interesse agrícola, e conhecer melhor a origem evolutiva da diversidade existente neste grupo é crucial para isto.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Testar a viabilidade da utilização de marcadores obtidos a partir de sequenciamento RAD para reconstrução da história filogenética de *Barbacenia*.

2.2 OBJETIVOS ESPECÍFICOS

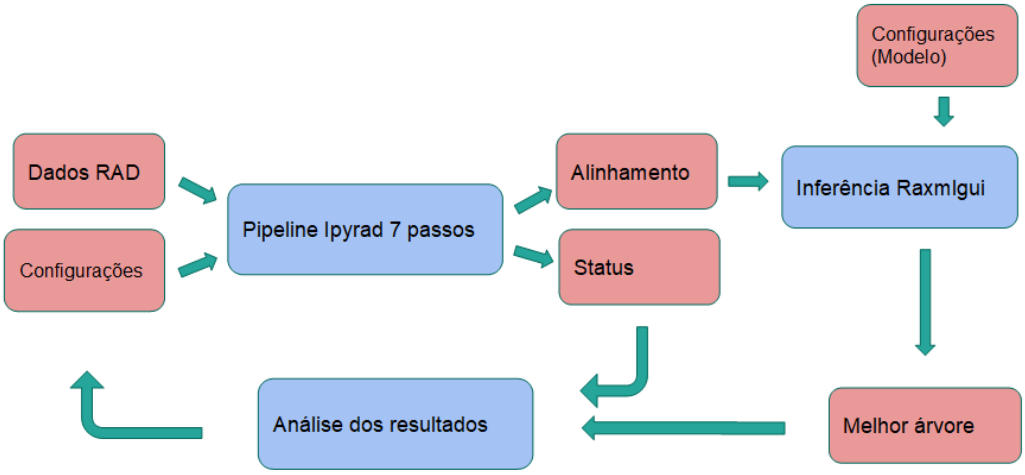
I. Inferir a árvore filogenética a partir do método de máxima verossimilhança;
II. Comparar a filogenia inferida com a filogenia atualmente disponível do grupo para avaliar a eficácia dos marcadores RAD na resolução de relações de parentesco que ainda não estão bem resolvidas.

3 MATERIAL E MÉTODOS

3.1 INTRODUÇÃO AO MÉTODO

Este trabalho consistiu essencialmente em utilizar dados de sequenciamento genético de diversas espécies e processá-los para inferir uma árvore filogenética através do uso de programas de computador especializados. Utilizei o programa ipyrad versão 0.7.28 para separar ou demultiplexar os dados, filtrar, mapear, alinhar, estimar heterozigose e gerar os arquivos de alinhamento no formato phylip. Em seguida, o alinhamento foi analisado no programa que compara os alinhamentos de cada táxon e tentam estimar a árvore mais provável de relação entre estes com base nas variações nos dados. Para isso utilizei o programa raxmlgui 1.5b1 (SILVESTRO; MICHALAK, 2012); disponível em: https://sourceforge.net/projects/raxmlgui/). Frequentemente, a extensão de arquivo que um programa oferece como output não é compatível com o input esperado pelo próximo programa na cadeia de análise, para isso utilizei o programa de manipulação de matrizes Mesquite 3.6 (MADDISON, 2018); disponível em: http://www.mesquiteproject.org), também utilizado para visualizar e editar árvores prontas. Com o resultado do processo, pude analisar parâmetros como a probabilidade de sustentação do nó e comprimento de ramo. Estes dados por sua vez podem ser utilizados para guiar e otimizar o alinhamento, na tentativa de melhorar o ajuste do modelo evolutivo utilizado. Finalmente, a melhor árvore emitida foi comparada com o conhecimento atual do gênero para verificar se os resultados são congruentes e se houve acréscimo de resolução filogenética com a utilização de marcadores RAD (Figura 1).

Figura 1 – Fluxograma da análise dos dados.



Fonte: Elaborada pelo autor (2019).

3.2 AMOSTRAGEM FILOGENÉTICA E SEQUÊNCIAS RAD UTILIZADAS

Os dados genômicos foram disponibilizados pela professora Dra. Suzana Alcantara (orientadora deste projeto) para análise. As sequências compreendem a amostragem de 33 indivíduos do gênero *Barbacenia*, representando 31 das 48 espécies (APÊNDICE A) amostradas na filogenia molecular disponível para o grupo (ALCANTARA; REE; MELLO-SILVA, 2018). As extrações de DNA foram feitas a partir do método CTAB no Pritzker Laboratory, em Chicago, e o DNA resultante preparado para envio para empresa que realiza o processo de sequenciamento RAD (Floragenex: www.floragenex.com). O método de sequenciamento utilizado foi o RadSeq (*Restricted site-associated DNA sequencing*) do tipo *single end*, onde diversas amostras são lidas na mesma placa, e a duplicação do material é feita apenas com um *primer* inicial. Neste trabalho, as sequências foram obtidas a partir da clivagem do DNA total das amostras com a enzima de restrição Pst1 (WALDER; WALDER; DONELSONG, 1984), ancorada a *barcodes* específicos para cada amostra, e submetidas a corrida multiplex em uma única linha de sequenciador Illumina GAIIx por 75 ciclos (gerando sequências *single-end* - chamadas aqui de *reads*). Esses dados foram disponibilizados em arquivos no formato de *output* Illumina FASTQ, e analisados para o alinhamento e posterior inferência da relação filogenética entre as espécies.

3.3 PROCESSAMENTO IPYRAD

O processamento dos dados possui muitos passos e felizmente, pude utilizar *pipelines* de processamento para tornar o processo mais ágil. O processamento dos dados inicia com a *pipeline* de *assemblage* e alinhamento chamada ipyrad (EATON, 2014, disponível em: https://github.com/dereneaton/ipyrad). Este programa foi desenvolvido em linguagem python especialmente para processar dados de sequenciamento RAD. O comando inicial gera um arquivo de parâmetros que são utilizados nos diversos passos da *pipeline* (ANEXO 1). Em seguida, o primeiro passo chamado demultiplex é executado para que o grande arquivo indexado fornecido pela empresa de sequenciamento seja descompactado em vários arquivos menores com os dados de cada terminal sequenciado. Então, um comando chamado *branching* (ANEXO 2) é utilizado para selecionar somente as espécies a serem utilizadas no alinhamento. Isso inclui todos os terminais disponíveis do gênero *Barbacenia*, mais um táxon próximo e sabidamente não incluso ao grupo para ser utilizado como *outgroup* nas análises posteriores. O processamento inicia filtrando os pequenos pedaços de leitura chamados *reads* com base na qualidade da leitura relatada pelo sequenciador. Após eliminar os *reads* de baixa qualidade o programa contabiliza o número de *reads* repetidos e tenta agrupá-los em grupos chamados *clusters*, que por sua vez são alinhados a *clusters* de outras entradas. Em seguida, o programa faz uma estimativa de heterozigose com base na ploidia configurada pelo usuário. Com isso um consenso é alcançado e os locais com um certo número mínimo de bases indeterminadas é filtrado. Finalmente, um último agrupamento é feito e alinhado entre as entradas dos terminais. Com o alinhamento terminado, o próximo passo é comparar os resultados obtidos para inferir a relação de parentesco entre os táxons sequenciados através de uma análise de inferência de máxima verossimilhança.

Finalmente, depois de alterar os parâmetros no arquivo, os demais passos da *pipeline* são executados. Diversos parâmetros foram alterados individualmente e alinhados na tentativa de encontrar a configuração que oferecesse melhor resolução. Parâmetros como número máximo de bases de baixa qualidade por *read* (param. 9) e porcentagem de semelhança para clusterização (param. 14) foram considerados. A variável que determina o número máximo de alelos por região (param. 18) foi alterado para avaliar ploidia. O número máximo de *mismatches* permitido em *barcodes* (param. 15) e a exigência para adaptadores de filtros (param. 16) também foram sondados. Uma variável que se mostrou importante foi o

comprimento mínimo de *reads* após filtragem (param. 17). Outra variável considerada foi o número máximo de snps por *locus* (param. 22). Os principais parâmetros foram listados e brevemente descritos no Quadro 1, onde também podemos ver quais destes parâmetros foram sondados através da modificação e análise do efeito desta modificação nos resultados. Alguns parâmetros não relevantes não foram mencionados.

Quadro 1 – Principais parâmetros da pipeline ipyrad.

Parâmetro	**Nome**	**Função**	**Padrão**	**Alterado**
7	*Datatype*	Determina o tipo de dado a ser sequenciado	rad	
8	*Restriction_overhang*	Retira sequências da extremidade da leitura inerentes do processo de sequenciamento	TGCAG	X
9	*max_low_qual_bases*	Número máximo de leituras de baixa qualidade por *read*	5	
10	*Phred_Qscore_offset*	Qualidade mínima para inclusão de uma leitura	33	X
11	*mindepth_statistical*	Profundidade estatística da seleção de bases	6	
12	*mindepth_majrule*	Profundidade estatística da seleção de bases por maioria absoluta	6	
13	*maxdepth*	Profundidade estatística máxima para clusterização	10000	
14	*clust_threshold*	Limite de similaridade para clusterização	0.85	X
15	*max_barcodes_mismatch*	Número máximo de diferenças entre o arquivo *barcode* e as leituras sequenciadas	0	X
16	*filter_adapters*	Filtragem de adaptadores Illumina	0	X
17	*filter_min_trim_len*	Comprimento mínimo para a remoção de adaptadores e filtragem	35	X
18	*max_alleles_consens*	Número máximo de alelos considerados	2	X
19	*max_Ns_consens*	Número máximo de frações não alinhadas permitidas no consenso	5,5	
20	*max_Hs_consens*	Número máximo de traços de bases heterozigotas permitidas no consenso	8,8	
21	*min_samples_locus*	Número mínimo de amostras por locus	4	
22	*max_SNPs_locus*	Número máximo de snps permitidos por locus	20,2	X
23	*max_Indels_locus*	Número máximo de indels permitidos por locus	8,8	
24	*max_shared_Hs_locus*	Número máximo de polimorfismos compartilhados em um locus	0.5	

Fonte: Elaborado pelo autor (2019).

O resultado de cada execução da *pipeline* são dois arquivos, um de alinhamento e um arquivo de texto com um relatório do processo, contendo informações como número de snps encontrados, pares de base sequenciados, etc. Um resumo dos dados é apresentado aqui e os detalhes completos listados no ANEXO 3, 4 e 5.

3.4 ANÁLISE DE INFERÊNCIA DE MÁXIMA VEROSIMILHANÇA COM RAXML

O arquivo phylip gerado pelo ipyrad pode ser utilizado diretamente como *input* no raxmlgui (SILVESTRO; MICHALAK, 2011). O arquivo é carregado no programa e é possível escolher o terminal a ser utilizado como *outgroup*, e o modelo para estimativa de mutação através das gerações. A definição de grupo externo (*outgroup*) é necessária em muitas análises filogenéticas, embora em outras o próprio algoritmo utilizado é capaz de enraizar a árvore filogenética utilizando como *outgroup* o grupo/terminal mais divergente em relação aos demais terminais analisados. A definição de um *outgroup* permite inferir quais homologias presumidas indicam relação genealógica e quais são somente caracteres plesiomórficos (MORRISON, 2012). Nas minhas análises, tentei criar alinhamentos com diferentes membros do gênero *Vellozia* (outro gênero de Velloziaceae diverso no Brasil) como *outgroup* na tentativa de obter melhor resolução.

O modelo de substituição de nucleotídeo que tem sido utilizado para análises filogenéticas baseadas em dados RAD é o GTRGAMMA + i, um modelo temporal geral reversível, em que a frequência de distribuição gamma varia entre os locais, que também considera sítios sem mutações ou invariáveis. Este é o modelo mais complexo disponível pelo programa utilizado, pois possui o maior número de parâmetros. O modelo GTR não só computa os parâmetros de alteração entre cada par de base (A,C,T,G), quanto a frequência estacionária de cada par de base em uma matriz. O parâmetro de alterabilidade permite considerar tendências de transição e transversão dos pares de bases. Junto a isso, a distribuição gamma permite considerar alterações nas frequências de mutação ao longo das regiões e dos terminais. Um parâmetro adicional é o cálculo de inferência de invariabilidade (BALDING; MOLTKE; MARIONI, 2019). O modelo é eficiente para a inferência filogenética em diversos grupos (BUCKLEY; CUNNINGHAM, 2002) e também na inferência de grupos de plantas com o mesmo tipo de sequenciamento (EATON et al., 2016).

O output desse programa são arquivos contendo árvores no formato *newick* com maiores valores de verossimilhança. Estes arquivos .tree são então visualizados e editados no Mesquite v. 3.6.

3.5 COMPARAÇÃO COM O CONHECIMENTO ATUAL DO GRUPO

Para comparação posterior, foram utilizados os dados de alinhamento usados por (ALCANTARA; REE; MELLO-SILVA, 2018), que foram obtidos através do sequenciamento do tipo Sanger de 4 marcadores: 1. trnL intron - trnL-F spacer (TABERLET ET AL., 1991); 2. atpB-rbcL spacer (CHIANG ET AL., 1998); 3. trnH-psbA spacer (SHAW ET AL., 2005) e 4. ITS rnDNA (SUN ET AL., 1994.) (APÊNDICE 1). Este alinhamento foi processado utilizando raxmlgui com os mesmos parâmetros utilizados para os dados RAD, e com o mesmo *outgroup*.

Em seguida, outra análise foi feita unindo as duas matrizes. A matriz RAD com 31 terminais e a matriz com os quatro marcadores Sanger contendo 48 terminais foram concatenadas e o programa raxmlgui foi utilizado mais uma vez para a inferência de máxima verosimilhança com o modelo GTRGAMMA + i.

4 RESULTADOS

Os melhores resultados obtidos para o alinhamento foram alcançados relaxando um pouco os parâmetros padrão do programa ipyrad, com exceção do parâmetro 17, que se refere ao comprimento mínimo de *reads* após filtragem (*trim*). Parâmetros como número máximo de bases de baixa qualidade por *read* (param. 9) e porcentagem de semelhança para clusterização (param. 14) foram relaxados mas estas análises geraram uma perda de resolução em relação às configurações padrão. A variável que determina o número máximo de alelos por região (param. 18) foi alterado para 3 e 4 na esperança de que diferenças em ploidia aumentassem a resolução da análise, mas isto também não ocorreu. O número máximo de *mismatches* permitido em *barcodes* (param. 15) foi aumentado para 1 e a exigência para adaptadores de filtros (param 16) também foi testada com todos os parâmetros possíveis (0, 1 e 2).

O parâmetro *restriction overhang* (param. 8) foi deixado em branco já que após análise dos dados crus, não foram encontradas sequências repetidas que podem ser inerentes ao processo de sequenciamento. A variável *clustering threshold* (param. 14) foi definida para 0.82. O número mínimo de pares de bases aceitáveis após filtragem (param. 17) foi aumentado até 50. Este valor, que como padrão é estabelecido em 35 pares de base, foi tanto relaxado quanto tornado mais restrito, e obtivemos os melhores valores de sustentação aumentando o comprimento mínimo dos *reads* - e portanto o rigor - para 50 pares de base. Outra variável que ajudou a aumentar resolução foi o aumento do número máximo de snps por locus (param. 22). O valor padrão que é 20 foi ajustado para 24 nos melhores resultados. A entrada definida como outgroup que nos permitiu alcançar melhor resolução foram os dados da espécie *Vellozia auriculata*.

Durante todas estas rodadas, apesar de os valores de sustentação terem variado, a estrutura geral da árvore sofreu poucas alterações, com a exceção de terminais como *B. delicatula* e *B. macrantha*, que se posicionaram em diferentes regiões dependendo dos parâmetros escolhidos para cada alinhamento, e de outros nós mais derivados que ofereceram menor resolução no geral. Aqui, serão comparados os dois melhores alinhamentos, obtidos com parâmetros idênticos (ANEXO 6) com exceção do parâmetro 17, sendo este 31 para o alinhamento 1 e 50 para o alinhamento 2. Os números totais de *loci* pré filtragem são 8904 e 8883 e ao final, 2735 e 2832 loci foram retidos respectivamente, sendo que a maioria dos dados filtrados foram retidos por não alcançar amostragem mínima para clusterização,

número máximo de alelos por *locus* e número máximo de *snps* por *read.* Outros critérios filtrados foram números máximo de indels, duplicatas inerentes ao processo de PCR e número máximo de heterozigose compartilhada.(Tabela 1).

Tabela 1 - Resultados gerais da filtragem dos dados.

	Alinhamento 1		**Alinhamento 2**	
	filtrados totais	**loci retidos**	**filtrados totais**	**loci retidos**
total de loci prefiltrados	8904	8904	8883	8883
filtrados por rm duplicados	375	8529	375	8508
filtrados por max de indels	865	7664	816	7692
filtrados por max snps	1081	7136	945	7263
filtrados por max heterosigose compartilhada	92	7106	88	7233
filtrados por amostra minima	4048	3153	4039	3288
filtrados por max alelos	1095	2735	1088	2832
filtrados totais por loci	2735	2735	2832	2832

Fonte: Elaborada pelo autor (2019).

Com isso foi possível obter dois alinhamentos com a maioria dos terminais apresentando mais de 1000 loci por alinhamento, com exceção de alguns terminais como *B. mantiqueirae*, *B. serracabralea*, *B. spectabilis* e do outgroup *V. auriculata* (ANEXO 3).

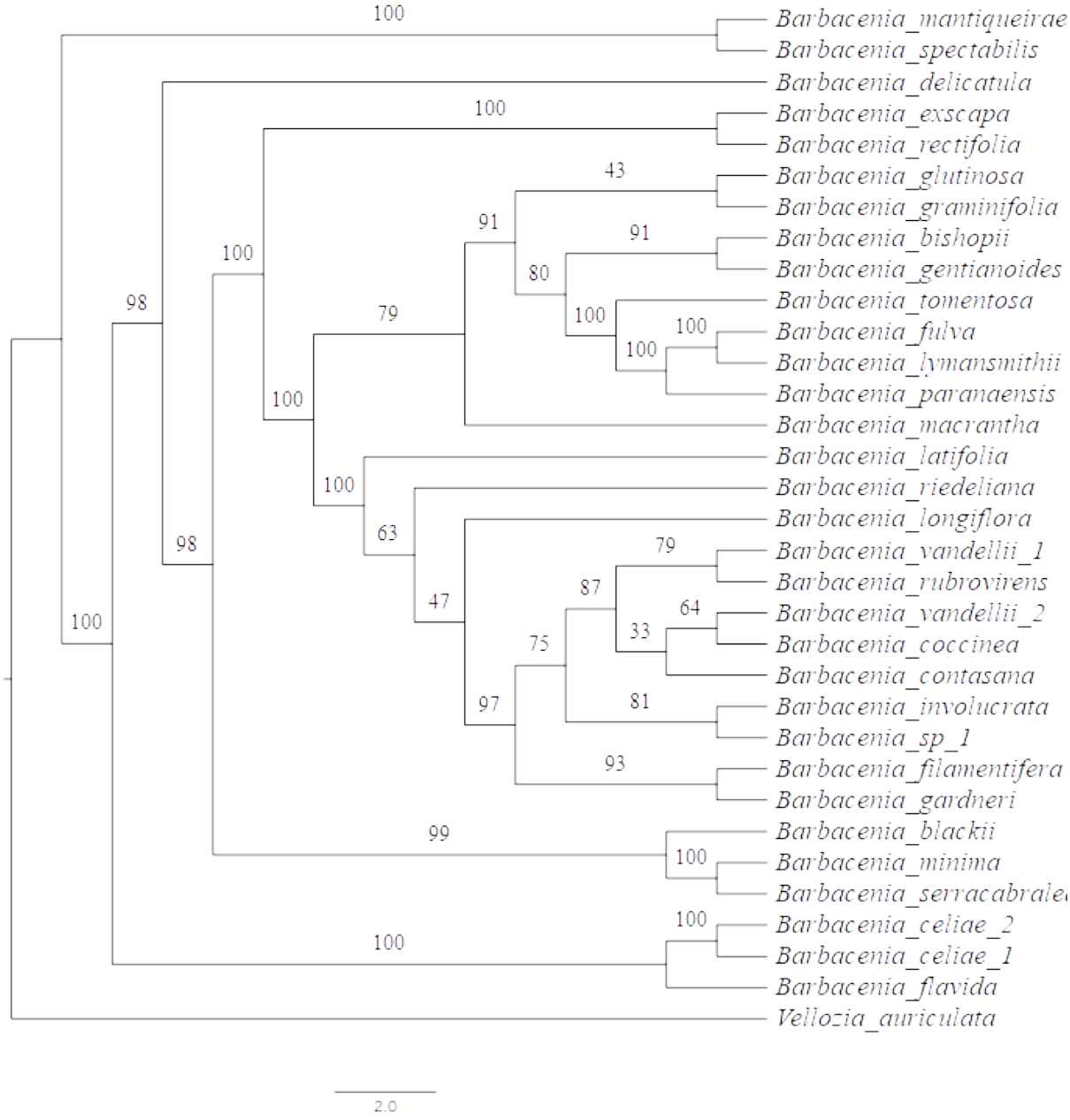
Ao analisar o número de loci informativos por região, foi possível notar que apesar do relaxamento do parâmetro 22, a grande maioria das regiões apresentou menos de 10 snps informativos por locus, sendo 20 o limite de snps informativos encontrados por locus (ANEXO 4). Com isso, o resultado final dos alinhamentos mostrou que o segundo alinhamento com uma restrição maior, delimitando o comprimento mínimo de pares de base para 50 recuperou mais loci informativos após o processamento dos dados (ANEXO 4).

Com os alinhamentos gerados, rodei uma análise de verossimilhança com o software raxmlgui. Após um bootstrap de 1000 vezes, a análise buscou a árvore com o melhor valor de verossimilhança e gerou um arquivo de árvore com os valores de sustentação dos ramos, junto com informações sobre a análise. Este processo foi feito para os dois alinhamentos discutidos neste texto e os valores inferidos pelo modelo foram alpha=0.300852 (alinhamento 1) e 0.323128 (alinhamento 2) para a estimativa gamma de variação. Os valores de estimativa de

loci invariantes foram 0.38 (alinhamento 1) e 0.39 (alinhamento 2). A variável comprimento da árvore resultou 0.738653 (alinhamento 1) e 0.776163 (alinhamento 2). As frequências de variação e frequências para cada par de base são mostradas no ANEXO 7.

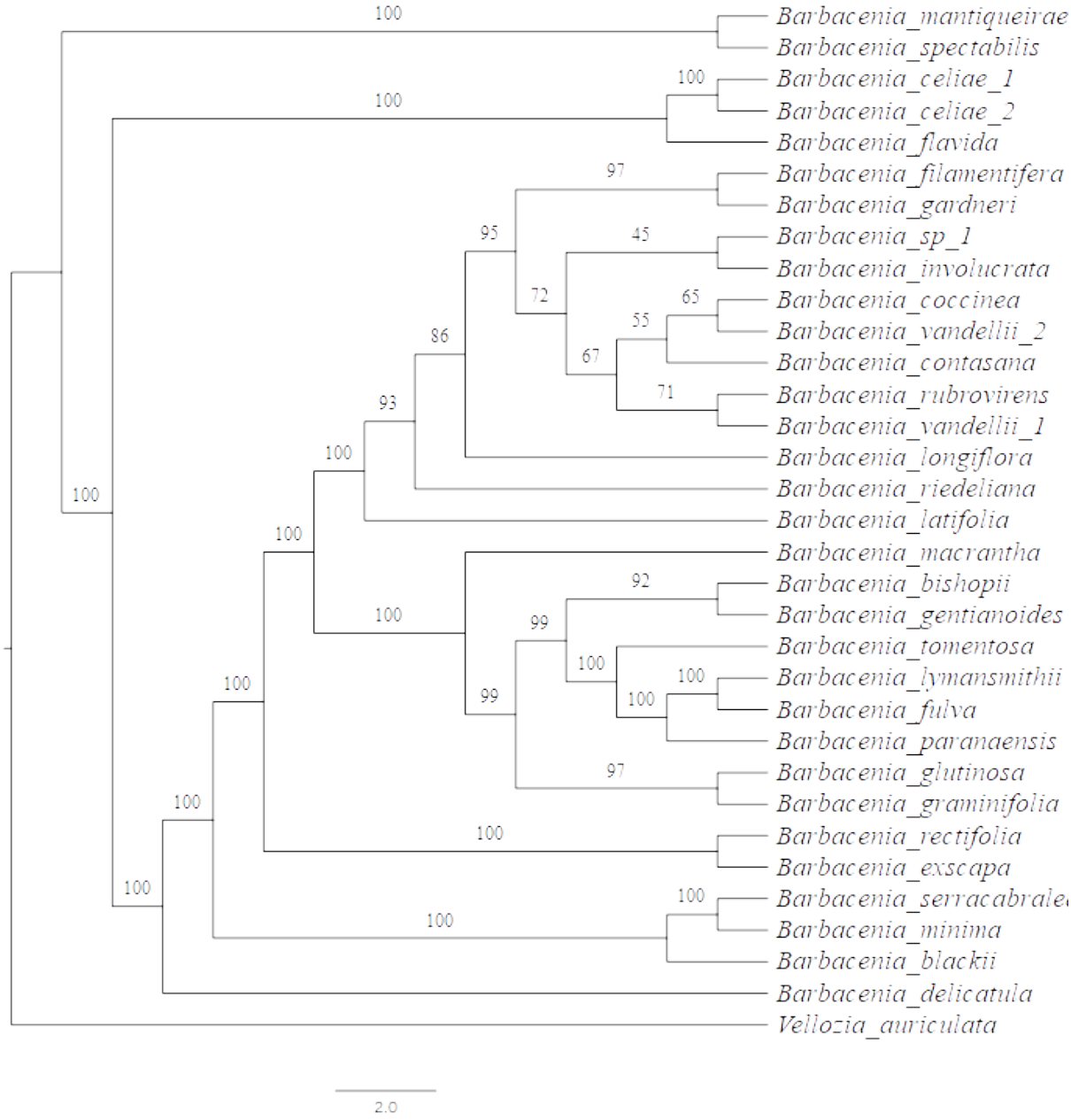
Para cada alinhamento, as mil árvores simuladas foram comparadas e as árvores com maior valor de verossimilhança foram geradas com o programa raxmlgui, a partir dos alinhamentos montados com os dois diferentes parâmetros de comprimento de *reads* (Figuras 1 e 2).

Figura 1 - Árvore inferida com alinhamento 1.



Fonte: Autor

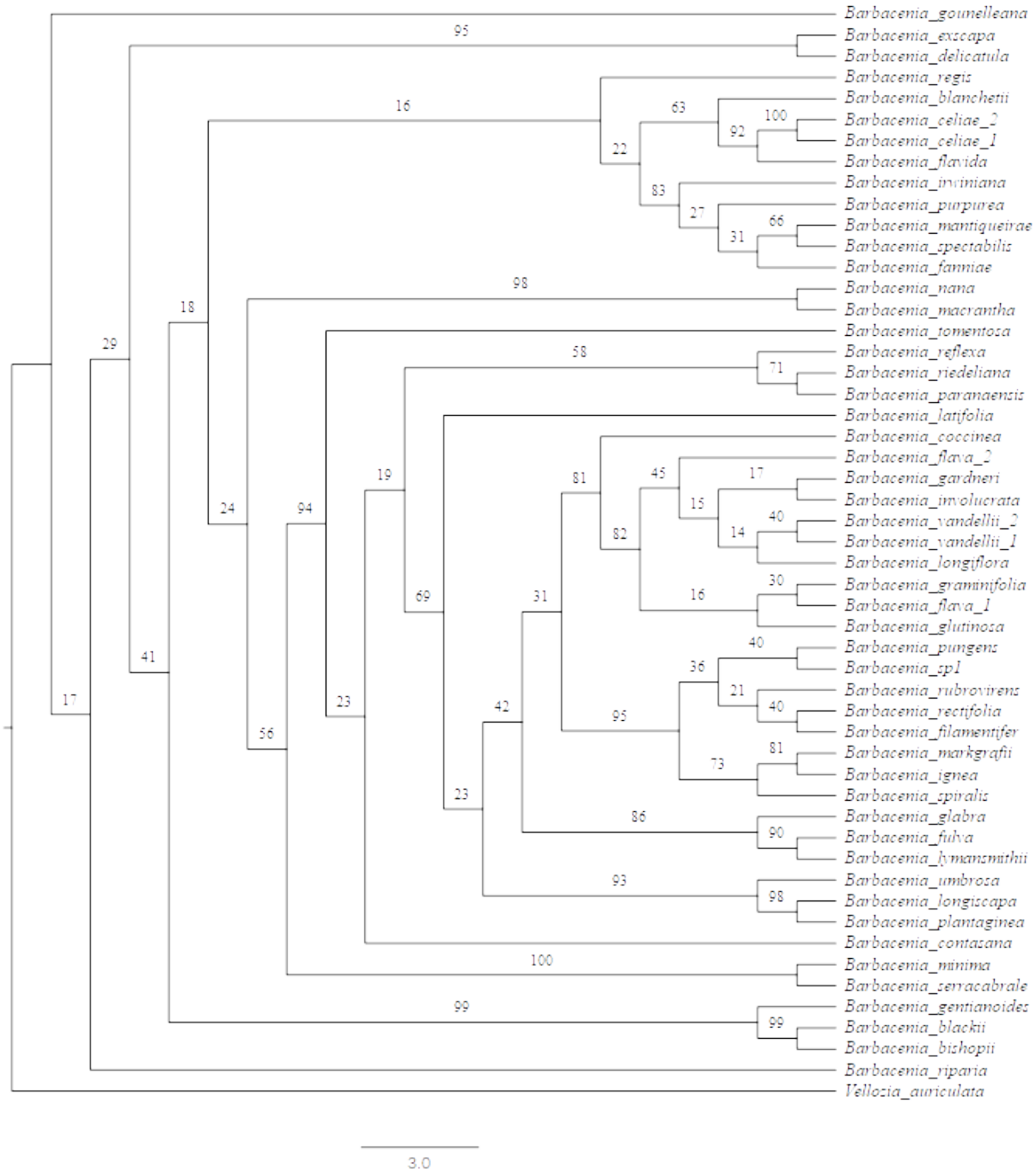
Figura 2 - Árvore inferida com alinhamento 2.



Fonte: Autor

Em seguida, fiz uma comparação inicial entre a árvore obtida com melhor resolução a partir dos marcadores RAD e a árvore gerada pela análise de máxima verosimilhança realizada com os marcadores (sequências) obtidos por sequenciamento do tipo Sanger para 48 espécies de *Barbacenia*, disponíveis em Alcantara, Ree e Mello-Silva (2018) (Figura 3).

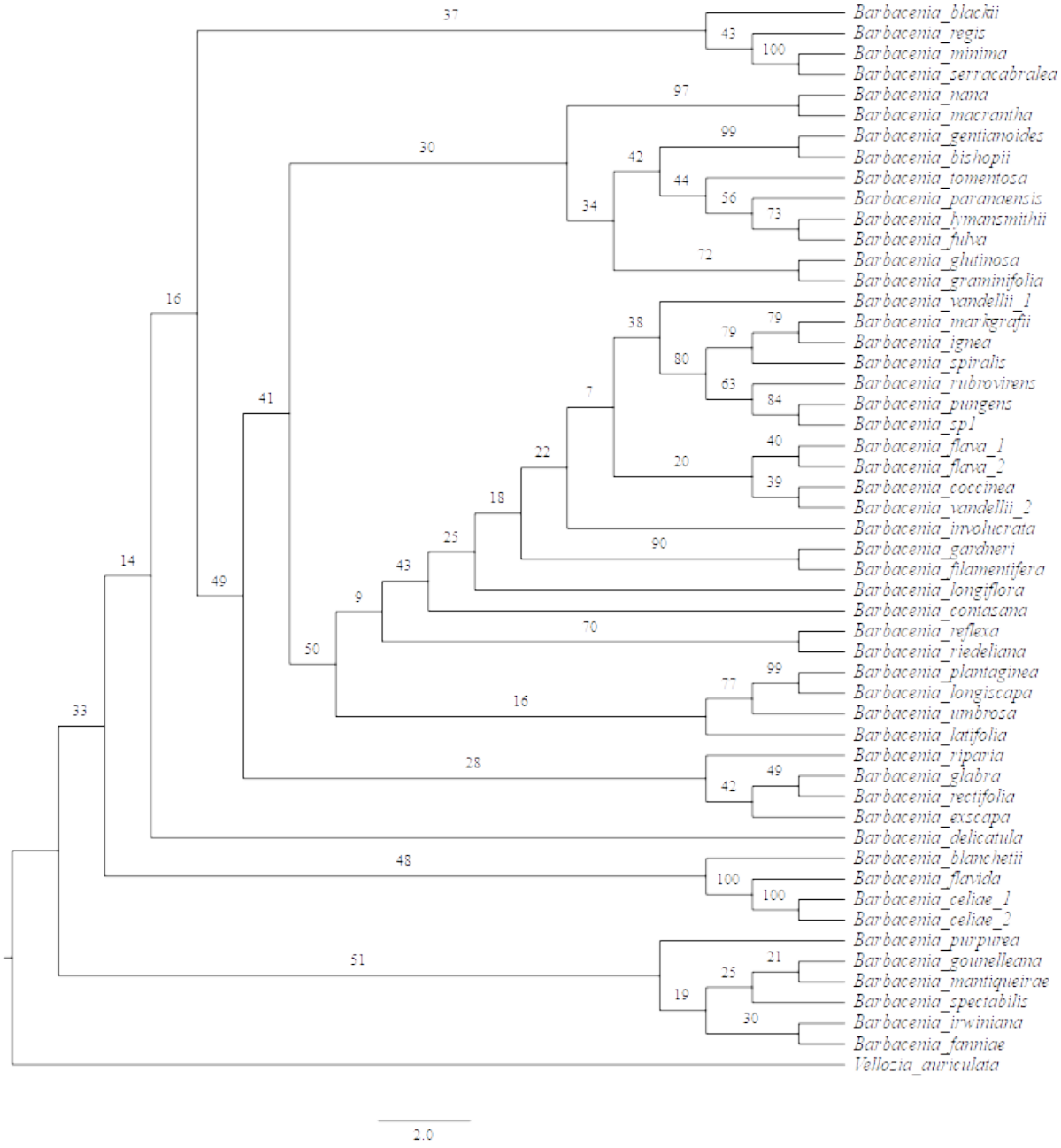
Figura 3 - Árvore inferida com os dados Sanger.



Fonte: Autor.

Finalmente, comparei visualmente as árvores anteriores com os resultados da análise feita com a matriz que combinou os dados RAD com os dados Sanger (Figura 4).

Figura 4 - Árvore inferida das matrizes RAD e Sanger combinadas.



Fonte: Autor

5 DISCUSSÃO

Com a necessidade de mapear sequências de aminoácidos de proteínas, que é um processo altamente repetitivo e estatístico, surgiu a bioinformática e logo no início já era evidente que seria útil para investigações de sequenciamento genético (HAGEN, 2000). Esta área de estudo da biologia trabalha a montagem de genomas a partir de pedaços menores lidos durante o sequenciamento. Isso inclui manipular as leituras para a remoção de primers, aplicação de filtros e união de sequências montadas através de extensa comparação entre as bases nitrogenadas (WIEL et al., 2010). Com o avanço destas tecnologias, a disponibilidade de dados de sequenciamento obtidos por diferentes métodos está cada vez mais acessível para os pesquisadores, principalmente para o objetivo de reconstrução de árvores filogenéticas. O surgimento de sequenciadores de segunda geração do tipo Illumina e a redução do custo de sequenciamento do tipo RAD é um fator importante (MILLER et al., 2007). Outro fator crucial é a criação de *pipelines* de processamento, que são *scripts* desenvolvidos para automatizar e conectar os vários passos deste processo, reduzindo o custo da manipulação destes dados em mão de obra e em tempo.

Logo no início das análises de máxima verossimilhança foi possível verificar que os marcadores sequenciados se mostram informativos para a inferência da filogenia em questão, especialmente no que diz respeito aos ramos mais basais. Mas foi somente com a utilização de modelos apropriados de mutação, juntamente com um alinhamento cuidadosamente calibrado que foi possível alcançar altos níveis de sustentação para o esqueleto geral da filogenia do grupo. O fato do aumento do comprimento mínimo de *reads* gerar uma maior resolução corrobora a hipótese de que métodos modernos de processamento de dados RAD são capazes de encontrar sequências ortólogas em meio aos dados crus de sequenciamento por digestão enzimática aleatória. Porém, mais investigação é necessária já que esta melhora pode ter sido dada por vários fatores, e essa é uma das principais críticas atuais ao método devido a distribuição desconhecida dos dados faltantes (EATON et al., 2016). Foi possível verificar que mesmo com as diferenças em sustentação, as duas árvores geradas com os melhores alinhamentos RAD apresentam exatamente a mesma topologia, e um esqueleto bem definido sustentando as relações de divergência mais profunda no gênero.

Uma comparação com a filogenia molecular disponível para a família (ALCANTARA; REE; MELLO-SILVA, 2018) indica que a filogenia obtida com marcadores

do tipo Sanger tem maior capacidade de resolver ramos mais recentes, apesar de não oferecer a mesma resolução em relações mais basais (ver Figura 3 acima). Isso contrasta com a inferência dos alinhamentos RAD, que mostraram bons resultados ao longo de toda a árvore, principalmente nos ramos mais basais (Figuras 1 e 2). Uma comparação visual foi inicialmente feita para buscar incongruências óbvias entre as árvores. Como a árvore RAD possui baixa sustentação nas relações entre os terminais e a árvore com marcadores Sanger possui pouca sustentação nos ramos mais basais, não foi possível verificar contradições claras. Com base na falta de resolução complementar desses marcadores, realizamos a análise combinada com a expectativa que a combinação desses conjuntos de dados forneceria uma melhoria significativa da filogenia do gênero. Apesar disso, a análise da matriz combinada apresentou pequena melhora de sustentação, o que é um indício de incongruência entre estes dados.

Para avaliar a causa da (aparente) incongruência entre as filogenias obtidas com os marcadores RAD e Sanger em *Barbacenia*, pretendo refazer o alinhamento RAD com base no genoma de referência de *Vellozia nivea*, que está sendo finalizado pelos pesquisadores Isabel Gerhardt e Ricardo Dante da Embrapa Informática Agropecuária (CNPTIA). Como esta é uma espécie que possuímos na nossa matriz RAD, isso maximizará as nossas chances de identificar sequências ortólogas durante o alinhamento. De acordo com esses colaboradores, o genoma completo de *Vellozia nivea* será disponibilizado para que possamos utilizá-lo nessas análises no mais tardar ao longo do mês de fevereiro, já que sua anotação está em fase de finalização. Além disso, análises específicas para avaliar a incongruência entre esses dados serão realizadas com o uso dos *softwares* concaterpillar (LEIGH et al., 2008; disponível em http://rogerlab.biochemistryandmolecularbiology.dal.ca/ccp.php), que é um método de clusterização hierárquica baseado em verossimilhança que identifica loci congruentes, e com o PAUP (Swofford, 2003;disponível em: http://phylosolutions.com/paup-test/), que é um software para análises filogenéticas que oferece uma análise de incongruência de diferença entre comprimentos de ramo. Essas análises serão realizadas visando incorporar as especificidades e diferenças dos dois conjuntos de dados e só não foram feitas pela limitação de tempo para a familiarização com os métodos de análise de congruência e devido ao tempo de análise computacional necessária. Outro ponto que merece atenção é que todas as análises dos dados Sanger foram feitas com as configurações padrão ao que refere as partições dos

arquivos de alinhamento. Estas partições podem ser utilizadas posteriormente para selecionar modelos evolutivos mais adequados para cada marcador utilizado.

Apesar da definição dos nós de divergência mais profunda em *Barbacenia*, nem todos os ramos puderam ser definidos com 100% de probabilidade. Isso pode ser resultado de radiações evolutivas, em que um grande número de eventos de especiação ocorrem em um intervalo cronológico muito pequeno. Nestes casos, não há acúmulo suficiente de mutações que permitam uma inferência mais precisa por meio de quaisquer métodos filogenéticos, o que é congruente com as análises de diversificação que apontam uma maior taxa de especiação para o grupo *Barbacenia* nos últimos 15 milhões de anos, em comparação com os outros gêneros da família (ALCANTARA; REE; MELLO-SILVA, 2018). Por outro lado, essa falta de resolução entre espécies proximamente aparentadas (i.e., nós terminais na filogenia) também pode ser resultado de especiação incipiente e/ou presença de fluxo gênico e introgressão entre as espécies, o que não pode ser descartado com base na distribuição de muitas dessas espécies que são simpátricas e/ou ocorrem em áreas próximas. No entanto, para identificar se as causas dessa baixa resolução deve-se a fatores biológicos como os citados ou é um artefato da obtenção de marcadores idênticos por similaridade convergente e não por descendência, é crucial a realização do alinhamento com base em um genoma de referência para avaliar quantos desses marcadores podem ser posicionados ao longo do genoma.

6 CONCLUSÃO

Com tudo pude concluir que a utilização do método de sequenciamento RAD, quando processados de maneira cuidadosa podem gerar filogenias com alta sustentação em nível infra-genérico. Este tipo de sequenciamento oferece uma riqueza de informação para o uso do pesquisador, e estes mesmos dados podem ser explorados de diversas formas que fogem do escopo deste trabalho. Apesar disso, a ausência do ganho em resolução ao unir as matrizes RAD e Sanger analisadas sugerem uma incongruência nos dados, que deve ser testada através de programas de teste de congruência antes da publicação destes dados em revista científica.

7 REFERÊNCIAS

AIDAR, S. T. et al. Desiccation tolerance in Pleurostima purpurea (Velloziaceae). **Plant Growth Regulation**, v. 62, n. 3, p.193-202, jun. 2010. Springer Science and Business Media LLC.

ALCANTARA, Suzana et al. Carbon assimilation and habitat segregation in resurrection plants: a comparison between desiccation- and non-desiccation-tolerant species of Neotropical Velloziaceae (Pandanales). **Functional Ecology**, v. 29, n. 12, p.1499-1512, maio 2015. Wiley.

ALCANTARA, Suzana; REE, Richard H; MELLO-SILVA, Renato. Accelerated diversification and functional trait evolution in Velloziaceae reveal new insights into the origins of the campos rupestres' exceptional floristic richness. **Annals Of Botany**, v. 122, n. 1, p.165-180, maio 2018. Oxford University Press (OUP).

BALDING, David; MOLTKE, Ida; MARIONI, John. **Handbook of Statistical Genomics: Two Volume Set**. 4. ed. Si: Wiley, 2019. 1135 p.

BUCKLEY, Thomas R.; CUNNINGHAM, Clifford W.. The Effects of Nucleotide Substitution Model Assumptions on Estimates of Nonparametric Bootstrap Support. **Molecular Biology And Evolution**, v. 19, n. 4, p.394-405, abr. 2002. Oxford University Press (OUP).

CHIANG, Tzen-Yuh; SCHAAL, Barbara; PENG, Ching-I. Universal primers for sequencing a noncoding spacer between the atpB and rbcL genes of chloroplast DNA.
Botanical Bulletin of Academia Sinica 39: 245–250. abr. 1998. Academia Sinica Taipei.

CHOI, Bokyung et al. Identifying genetic markers for a range of phylogenetic utility–From species to family level. **Plos One**, v. 14, n. 8, ago. 2019. Public Library of Science (PLoS).

EATON, Deren A. R. et al. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. Systematic Biology, p.399-412, out. 2016. Oxford University Press (OUP).

EATON, Deren A. R.. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics, v. 30, n. 13, p.1844-1849, mar. 2014. Oxford University Press (OUP).

EATON, Deren A. R.; REE, Richard H.. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis. **Systematic Biology**, v. 62, n. 5, p.689-706, 14 jun. 2013. Oxford University Press (OUP).

FELSENSTEIN, Joseph. **Inferring Phylogenies**. 2. ed. Si: Sinauer Associates, 2003. 580 p.

FITZ-GIBBON, Sorel et al. Phylogenomic inferences from reference-mapped and de novo assembled short-read sequence data using RADseq sequencing of California white oaks

(Quercus section Quercus). **Genome**, v. 60, n. 9, p.743-755, set. 2017. Canadian Science Publishing.

GAFF, D. F.. Desiccation tolerant vascular plants of southern Africa. **Oecologia**, v. 31, n. 1, p.95-109, 1977. Springer Nature.

HAGEN, Joel B. The origins of bioinformatics. **Nature Reviews Genetics**, v. 1, n. 3, p.231-236, dez. 2000. Springer Nature.

HOU, Yan et al. Thousands of RAD-seq Loci Fully Resolve the Phylogeny of the Highly Disjunct Arctic-Alpine Genus Diapensia (Diapensiaceae). **Plos One**, v. 10, n. 10, out. 2015. Public Library of Science (PLoS).

KERSTEN, Roland D.; WENG, Jing-ke. Gene-guided discovery and engineering of branched cyclic peptides in plants. **Proceedings Of The National Academy Of Sciences**, v. 115, n. 46, p.10961-10969, 29 out. 2018. Proceedings of the National Academy of Sciences.

LEIGH, Jessica W. et al. Testing Congruence in Phylogenomic Analysis. **Systematic Biology**, v. 57, n. 1, p.104-115, fev. 2008. Oxford University Press (OUP).

LUSCOMBE, N. M.; GREENBAUM, D.; GERSTEIN, M.. What is Bioinformatics? A Proposed Definition and Overview of the Field. **Methods Of Information In Medicine**, v. 40, n. 04, p.346-358, 2001. Georg Thieme Verlag KG.

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

MELLO-SILVA, Renato et al. Five vicarious genera from Gondwana: the Velloziaceae as shown by molecules and morphology. **Annals Of Botany**, v. 108, n. 1, p.87-102, jan. 2011. Oxford University Press (OUP).

MILLER, M. R. et al. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. **Genome Research**, v. 17, n. 2, p.240-248, jan. 2007. Cold Spring Harbor Laboratory.

MORRISON, David A.. Phylogenetics: The Theory and Practice of Phylogenetic Systematics, 2nd edition.—E. O. Wiley and Bruce S. Lieberman. **Systematic Biology**, v. 61, n. 6, p.1087-1088, 13 ago. 2012. Oxford University Press (OUP).

RUBIN, Benjamin E. R.; REE, Richard H.; MOREAU, Corrie S. Inferring Phylogenies from RAD Sequence Data. **Plos One**, v. 7, n. 4, abr. 2012. Public Library of Science (PLoS).

SATLER, Jordan D. et al. Inferring processes of coevolutionary diversification in a community of Panamanian strangler figs and associated pollinating wasps*. **Evolution**, v. 73, n. 11, p.2295-2311, ago. 2019. Wiley.

SHAW, Joey; LICKEY, Edgar; BECK, John et al. The tortoise and the hare II: Relative

utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. **American Journal of Botany** 92: 142–166. 2005. Botanical Society of America

SILVESTRO, Daniele; MICHALAK, Ingo. RaxmlGUI: a graphical front-end for RAxML. **Organisms Diversity & Evolution**, v. 12, n. 4, p.335-337, set. 2011. Springer Science and Business Media LLC.

SUN, Y; SKINNER, D; LIANG, G; HULBERT, S. Phylogenetic analysis of Sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. **Theoretical Application Genetics** 89: 26–32. set. 1994.

SWOFFORD, D.L. (2003) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. **Sinauer Associates**, Sunderland.

TABERLET, Pierre; GIELLY, Ludovic; PAUTOU, Guy; BOUVET, Jean. Universal primers for amplification of three non-coding regions of chloroplast DNA. **Plant Molecular Biology** 17: 1105–1109. Nov. 1991.

UNDA, F. Introduction to phylogenetics. The science creative quarterly. 2005. Disponível em: <www.scq.ubc.ca/introduction-to-phylogenetics/>. Acesso em 06 de nov. 2018.

WALDER, Roxanne Y.; WALDER, Joseph A.; DONELSONG, John E. The organization and complete nucleotide sequence of the PstI restriction-modification system. **The Journal Of Biological Chemistry**, v. 259, n. 12, p.8015-8026, jun. 1984.

WIEL, M. A. van de et al. Preprocessing and downstream analysis of microarray DNA copy number profiles. **Briefings In Bioinformatics**, v. 12, n. 1, p.10-21, fev. 2010. Oxford University Press (OUP).

8 APÊNDICE E ANEXOS

APÊNDICE A – Informações sobre a amostragem das espécies estudadas nesse trabalho.

As espécies que contam com dados de RadSeq são indicadas com x na última coluna. As demais espécies foram amostradas para reconstrução filogenética com base em marcadores do tipo Sanger (ALCANTARA; REE; MELLO-SILVA, 2018).

Quadro 2. - Informações de amostragem de espécies coletadas.

Species	**Voucher and herbaria**	**RadSeq**
Barbacenia bishopii L.B.Sm.	CFSC11228 (SPF)	X
Barbacenia blackii L.B.Sm.	Mello-Silva 1103 (SPF)	X
Barbacenia blanchetii Goethart & Henrard	Mello-Silva 2582 (SPF)	
Barbacenia celiae Maguire	Rodrigues s.n. (SPF); Rando 1171 (SPF)	X
Barbacenia coccinea Mart. ex Schult. & Schult.f.	Trovó 380 (SPF)	X
Barbacenia contasana L.B.Sm. & Ayensu	Mello-Silva 2136 (SPF)	X
Barbacenia delicatula L.B.Sm. & Ayensu	Mello-Silva 2668 (SPF)	X
Barbacenia exscapa Mart.	Mello-Silva 3466 (SPF)	X
Barbacenia fanniae (N.L.Menezes) Mello-Silva	Mello-Silva 2643 (RB, SPF)	
Barbacenia filamentifera L.B.Sm. & Ayensu	Mello-Silva 2555 (SPF); Mello-Silva 3404 (SPF)	X
Barbacenia flava Mart. ex Schult. & Schult.f.	Mello-Silva 2662 (SPF); Alcantara 195 (SPF)	

Barbacenia flavida Goethart & Henrard	Alcantara 160 (SPF)	X
Barbacenia fulva Goethart & Henrard	Mello-Silva 1675 (SPF)	X
Barbacenia gardneri Seub.	Mello-Silva 3414 (SPF)	X
Barbacenia gentianoides Taub. ex Goethart & Henrard	Mello-Silva 2370 (SPF)	X
Barbacenia glabra Goethart & Henrard	Trovó 393 (SPF)	
Barbacenia glutinosa Goethart & Henrard	Mello-Silva 3396 (SPF)	X
Barbacenia gounelleana Beauverd	Mello-Silva 2849 (SPF)	
Barbacenia graminifolia L.B.Sm.	Lovo 442	X
Barbacenia ignea Mart. ex Schult. & Schult.f.	Mello-Silva 2554 (B, K, RB, SPF, US)	
Barbacenia involucrata L.B.Sm.	Mello-Silva 3402 (SPF)	X
Barbacenia irwiniana L.B.Sm.	Trovó 252 (SPF)	
Barbacenia latifolia L.B.Sm. & Ayensu	Mello-Silva 1509 (SPF); Mello-Silva 2473 (SPF)	X
Barbacenia longiflora Mart.	Mello-Silva 3427 (SPF)	X
Barbacenia longiscapa Goethart & Henrard	Mello-Silva 2553 (K, SPF)	
Barbacenia lymansmithii Mello-Silva & N.L.Menezes	Farinaccio 139 (SPF)	X
Barbacenia macrantha Lem.	Mello-Silva 2900 (SPF)	X
Barbacenia mantiqueirae Goethart & Henrard	Trovó 323 (SPF)	X

Barbacenia markgrafii Schulze-Menz	Mello-Silva 1504 (BHCB, K, NY, SPF, W)	
Barbacenia minima L.B.Sm. & Ayensu	Mello-Silva 3435 (SPF)	X
Barbacenia nana L.B.Sm. & Ayensu	Miliken 4295 (K, SPF)	
Barbacenia paranaensis L.B.Sm.	Mello-Silva 3586 (SPF)	X
Barbacenia plantaginea L.B.Sm.	Salatino CFCR11901 (K, SPF)	
Barbacenia pungens (N.L.Menezes & Semir) Mello-Silva	Mello-Silva 319 (SPF)	
Barbacenia purpurea Hook.	Menezes 511 (SPF)	
Barbacenia rectifolia L.B.Sm. & Ayensu	Trovó 409 (SPF)	X
Barbacenia reflexa L.B.Sm. & Ayensu	Mello-Silva CFCR10793 (F, SPF)	
Barbacenia regis L.B.Sm.	Mello-Silva 2570 (SPF)	
Barbacenia riedeliana Goethart & Henrard	Mello-Silva 3565 (SPF)	X
Barbacenia riparia (N.L.Menezes & Mello-Silva) Mello-Silva	Menezes 1167 (SPF)	
Barbacenia rubrovirens Mart.	Mello-Silva 3405 (SPF)	
Barbacenia serracabralea Mello-Silva	Mello-Silva 2505 (B, BHCB, K, L, M, MBM, MO, NY, RB, SP, SPF, US); Mello-Silva 3439 (SPF)	X
Barbacenia sp. 1	Alcantara 118 (SPF)	X
Barbacenia spectabilis L.B.Sm. & Ayensu	Menezes 887 (SPF)	X

Barbacenia spiralis L.B.Sm. & Ayensu	Mello-Silva 2548 (SPF)	
Barbacenia tomentosa Mart.	Mello-Silva 1600 (K, MBM, RB, SP, SPF, UB, W); Mello-Silva 2929 (SPF); Prado 2166 (SP, SPF)	X
Barbacenia umbrosa L.B.Sm. & Ayensu	Mello-Silva CFCR9658 (F, K, MBM, RB, SPF)	
Barbacenia vandellii Pohl ex Seub.	Mello-Silva 3411 (SPF); Alcantara 200 (SPF)	X
Vellozia auriculata Mello-Silva & N.L.Menezes	Mello-Silva 3469 (SPF)	X

Fonte: modificado pelo autor a partir de ALCANTARA; REE; MELLO-SILVA, 2018.

ANEXO 1. Arquivo de parâmetros gerados na pipeline ipyrad, em formato .txt.

------- ipyrad params file (v.0.7.28)---

barbaceniaoutgvaurifinal ## [0] [assembly_name]: Assembly name. Used to name output directories for assembly steps

/ipyradprocessing/finalruns ## [1] [project_dir]: Project dir (made in curdir if not present)

/floragenex/UO_C291_1.gz ## [2] [raw_fastq_path]: Location of raw non-demultiplexed fastq files

/floragenex/multibarcodes.txt ## [3] [barcodes_path]: Location of barcodes file

[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

rad ## [7] [datatype]: Datatype (see docs): rad, gbs, ddrad, etc.

TGCAG, ## [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)

6 ## [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

0 ## [16] [filter_adapters]: Filter for adapters/primers (1 or 2=stricter)

35 ## [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)

8, 8 ## [20] [max_Hs_consens]: Max Hs (heterozygotes) in consensus (R1, R2)

4 ## [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)

p, s, v ## [27] [output_formats]: Output formats (see docs)

[28] [pop_assign_file]: Path to population assignment file

ANEXO 2. Comandos utilizados.

ipyrad -n test

ipyrad -p params-teste.txt -s 1

ipyrad -p params-barbrun.txt -b barbaceniaoutgvaurifinal3 Barbacenia_blackii Barbacenia_filamentifera Barbacenia_gardneri Barbacenia_glutinosa Barbacenia_longiflora Barbacenia_macrantha Barbacenia_minima Barbacenia_rubrovirens Barbacenia_spectabilis Barbacenia_vandellii_1 Barbacenia_coccinea Barbacenia_contasana Barbacenia_fulva Barbacenia_latifolia Barbacenia_mantiqueirae Barbacenia_rectifolia Barbacenia_serracabralea Barbacenia_tomentosa Barbacenia_exscapa Barbacenia_gentianoides Barbacenia_lymansmithii Barbacenia_riedeliana Barbacenia_celiae_2 Barbacenia_flavida Vellozia_auriculata Barbacenia_bishopii Barbacenia_delicatula Barbacenia_paranaensis Barbacenia_graminifolia Barbacenia_sp_1 Barbacenia_vandellii_2 Barbacenia_celiae_1 Barbacenia_involucrata Vellozia_auriculata

ipyrad -p params-barbaceniaoutgvaurifinal3.txt -s 234567

ANEXO 3. Cobertura em pares de base para cada terminal após alinhamento.

alinhamento 1

Barbacenia_bishopii	1190
Barbacenia_blackii	1061
*Barbacenia_celiae*_1	999
*Barbacenia_celiae*_2	1002
Barbacenia_coccinea	1369
Barbacenia_contasana	1137
Barbacenia_delicatula	1010
Barbacenia_exscapa	1095
Barbacenia_filamentifera	1458
Barbacenia_flavida	1074
Barbacenia_fulva	1152
Barbacenia_gardneri	1301
Barbacenia_gentianoides	1168
Barbacenia_glutinosa	1217
Barbacenia_graminifolia	1336
Barbacenia_involucrata	1362
Barbacenia_latifolia	1191
Barbacenia_longiflora	1485
Barbacenia_lymansmithii	1143
Barbacenia_macrantha	1403
Barbacenia_mantiqueirae	836
Barbacenia_minima	1114
Barbacenia_paranaensis	1073
Barbacenia_rectifolia	1140
Barbacenia_riedeliana	1275
Barbacenia_rubrovirens	1393
Barbacenia_serracabralea	913
*Barbacenia*_sp_1	1357

Barbacenia_spectabilis	828
Barbacenia_tomentosa	1236
*Barbacenia_vandellii*_1	1410
*Barbacenia_vandellii*_2	1077
Vellozia_auriculata	239

alinhamento 2

Barbacenia_bishopii	1261
Barbacenia_blackii	1123
*Barbacenia_celiae*_1	1053
*Barbacenia_celiae*_2	1058
Barbacenia_coccinea	1443
Barbacenia_contasana	1193
Barbacenia_delicatula	1062
Barbacenia_exscapa	1165
Barbacenia_filamentifera	1530
Barbacenia_flavida	1136
Barbacenia_fulva	1218
Barbacenia_gardneri	1371
Barbacenia_gentianoides	1239
Barbacenia_glutinosa	1289
Barbacenia_graminifolia	1399
Barbacenia_involucrata	1433
Barbacenia_latifolia	1264
Barbacenia_longiflora	1556
Barbacenia_lymansmithii	1215
Barbacenia_macrantha	1472
Barbacenia_mantiqueirae	875
Barbacenia_minima	1180
Barbacenia_paranaensis	1137
Barbacenia_rectifolia	1210
Barbacenia_riedeliana	1348

Barbacenia_rubrovirens	1463
Barbacenia_serracabralea	960
*Barbacenia*_sp_1	1426
Barbacenia_spectabilis	858
Barbacenia_tomentosa	1298
*Barbacenia_vandellii*_1	1476
*Barbacenia_vandellii*_2	1132
Vellozia_auriculata	251

ANEXO 4. Relatório de alinhamento listando o número de sítios variáveis, a somatória de sítios variáveis, o número de polimorfismos informativos por sítio e a somatória total do número de polimorfismos informativos por sítio.

Alinhamento 1.

	var	sum_var	pis	sum_pis
0	111	0	500	0
1	81	81	335	335
2	110	301	316	967
3	117	652	262	1753
4	121	1136	242	2721
5	125	1761	204	3741
6	148	2649	174	4785
7	121	3496	181	6052
8	119	4448	146	7220
9	129	5609	132	8408
10	136	6969	83	9238
11	133	8432	58	9876
12	133	10028	41	10368
13	120	11588	25	10693
14	118	13240	16	10917
15	122	15070	8	11037
16	108	16798	7	11149
17	101	18515	1	11166
18	113	20549	0	11166
19	107	22582	3	11223
20	86	24302	1	11243
21	66	25688	0	11243
22	73	27294	0	11243
23	81	29157	0	11243
24	56	30501	0	11243

Alinhamento 2

	var	sum_var	pis	sum_pis
0	120	0	523	0
1	82	82	352	352
2	122	326	338	1028
3	129	713	267	1829
4	127	1221	251	2833
5	131	1876	216	3913
6	154	2800	195	5083
7	126	3682	173	6294
8	141	4810	157	7550
9	135	6025	124	8666
10	149	7515	80	9466
11	132	8967	60	10126
12	152	10791	32	10510
13	114	12273	31	10913
14	126	14037	14	11109
15	117	15792	7	11214
16	101	17408	6	11310
17	124	19516	5	11395
18	108	21460	0	11395
19	89	23151	1	11414
20	85	24851	0	11414
21	73	26384	0	11414
22	68	27880	0	11414
23	61	29283	0	11414
24	66	30867	0	11414

ANEXO 5. Relatório de alinhamento listando detalhes do processo por espécie. Os dados em ordem são: Estado, referente ao passo alcançado na progressão da pipeline, *reads* crus, *reads* passados por filtros, número total de *clusters*, número de *clusters* após filtragem estatística, estimativa de heterozigose, estimativa de erro, *reads* inclusos no consenso e finalmente o número de *loci* recuperado no final da montagem.

Alinhamento 1

	state	reads_raw	reads_passed_filter	clusters_total	clusters_hidepth	hetero_est	error_est	reads_consens	loci_in_assembly
Barbacenia_bishopii	7	799454	799100	6313	3821	0.010238	0.002611	3713	1190
Barbacenia_blackii	7	1085026	1084560	7702	4105	0.013164	0.002229	3961	1061
Barbacenia_celiae_1	7	1635709	1634986	8026	3380	0.005841	0.001245	3276	999
Barbacenia_celiae_2	7	2749808	2748573	10306	3142	0.009736	0.000420	2993	1002
Barbacenia_coccinea	7	1133361	1132855	7182	3661	0.013991	0.002287	3534	1369
Barbacenia_contasana	7	302322	302197	4215	2895	0.009342	0.002665	2813	1137
Barbacenia_delicatula	7	1753109	1752305	10710	4688	0.005482	0.001534	4546	1010
Barbacenia_exscapa	7	2160590	2159597	12631	4820	0.014682	0.001331	4555	1095
Barbacenia_filamentifera	7	1047756	1047298	8162	4423	0.013512	0.002866	4267	1458
Barbacenia_flavida	7	1845689	1844903	11251	5053	0.006027	0.001431	4941	1074
Barbacenia_fulva	7	675376	675083	5161	3165	0.008954	0.002699	3069	1152
Barbacenia_gardneri	7	1048287	1047829	7585	4058	0.012292	0.002486	3924	1301
Barbacenia_gentianoides	7	518333	518075	7581	4933	0.008686	0.003025	4807	1168
Barbacenia_glutinosa	7	1070658	1070207	8123	4481	0.010104	0.002327	4356	1217
Barbacenia_graminifolia	7	1259367	1258850	7425	3707	0.015122	0.002393	3554	1336
Barbacenia_involucrata	7	1029502	1029045	6760	3574	0.011949	0.002447	3462	1362
Barbacenia_latifolia	7	431050	430856	5678	3695	0.010355	0.002958	3581	1191
Barbacenia_longiflora	7	936330	935898	8640	5073	0.012225	0.002847	4918	1485
Barbacenia_lymansmithii	7	724242	723929	6131	3753	0.008020	0.002320	3650	1143

Barbacenia_macrantha	7	1091234	1090736	8560	4820	0.012188	0.002603	4661	1403
Barbacenia_mantiqueirae	7	2331321	2330300	11029	4095	0.006852	0.000925	3968	836
Barbacenia_minima	7	1031695	1031249	7158	3741	0.014013	0.002162	3602	1114
Barbacenia_paranaensis	7	368637	368485	4509	3130	0.006135	0.002642	3066	1073
Barbacenia_rectifolia	7	1402331	1401696	8966	4176	0.017393	0.002080	3920	1140
Barbacenia_riedeliana	7	949645	949225	7724	4492	0.010149	0.002156	4363	1275
Barbacenia_rubrovirens	7	1034417	1033971	6885	3610	0.014103	0.002653	3482	1393
Barbacenia_serracabralea	7	264324	264222	4152	2753	0.007427	0.002907	2685	913
Barbacenia_sp_1	7	1128310	1127825	7102	3679	0.013511	0.002358	3560	1357
Barbacenia_spectabilis	7	4694403	4692306	17844	5250	0.009101	0.000464	5029	828
Barbacenia_tomentosa	7	619710	619403	5946	3613	0.013062	0.002874	3480	1236
Barbacenia_vandellii_1	7	948595	948156	7794	4390	0.011923	0.002611	4244	1410
Barbacenia_vandellii_2	7	254909	254796	3786	2713	0.008727	0.002457	2664	1077
Vellozia_auriculata	7	1701689	1700919	9388	4017	0.013078	0.001774	3838	239

Alinhamento 2

	state	reads_raw	reads_passed_filter	clusters_total	clusters_hidepth	hetero_est	error_est	reads_consens	loci_in_assembly
Barbacenia_bishopii	7	799454	799100	6313	3821	0.010520	0.002701	3712	1261
Barbacenia_blackii	7	1085026	1084560	7702	4105	0.013441	0.002298	3961	1123
Barbacenia_celiae_1	7	1635709	1634986	8026	3380	0.006117	0.001292	3276	1053
Barbacenia_celiae_2	7	2749808	2748573	10306	3142	0.010296	0.000436	2993	1058
Barbacenia_coccinea	7	1133361	1132855	7182	3661	0.014399	0.002356	3535	1443
Barbacenia_contasana	7	302322	302197	4215	2895	0.009568	0.002743	2813	1193
Barbacenia_delicatula	7	1753109	1752305	10710	4688	0.005748	0.001592	4546	1062
Barbacenia_exscapa	7	2160590	2159597	12631	4820	0.015179	0.001380	4555	1165
Barbacenia_filamentifera	7	1047756	1047298	8162	4423	0.013724	0.002955	4268	1530

Barbacenia_flavida	7	1845689	1844903	11251	5053	0.006159	0.001480	4942	1136
Barbacenia_fulva	7	675376	675083	5161	3165	0.008965	0.002789	3069	1218
Barbacenia_gardneri	7	1048287	1047829	7585	4058	0.012512	0.002568	3924	1371
Barbacenia_gentianoides	7	518333	518075	7581	4933	0.008292	0.003135	4807	1239
Barbacenia_glutinosa	7	1070658	1070207	8123	4481	0.010298	0.002415	4356	1289
Barbacenia_graminifolia	7	1259367	1258850	7425	3707	0.015371	0.002473	3554	1399
Barbacenia_involucrata	7	1029502	1029045	6760	3574	0.012121	0.002530	3462	1433
Barbacenia_latifolia	7	431050	430856	5678	3695	0.010492	0.003056	3581	1264
Barbacenia_longiflora	7	936330	935898	8640	5073	0.012316	0.002934	4918	1556
Barbacenia_lymansmithii	7	724242	723929	6131	3753	0.008447	0.002391	3650	1215
Barbacenia_macrantha	7	1091234	1090736	8560	4820	0.012350	0.002685	4661	1472
Barbacenia_mantiqueirae	7	2331321	2330300	11029	4095	0.007115	0.000962	3968	875
Barbacenia_minima	7	1031695	1031249	7158	3741	0.014294	0.002233	3602	1180
Barbacenia_paranaensis	7	368637	368485	4509	3130	0.006220	0.002776	3066	1137
Barbacenia_rectifolia	7	1402331	1401696	8966	4176	0.017650	0.002148	3921	1210
Barbacenia_riedeliana	7	949645	949225	7724	4492	0.010554	0.002222	4363	1348
Barbacenia_rubrovirens	7	1034417	1033971	6885	3610	0.014265	0.002742	3481	1463
Barbacenia_serracabralea	7	264324	264222	4152	2753	0.007201	0.003007	2685	960
Barbacenia_sp_1	7	1128310	1127825	7102	3679	0.013620	0.002439	3560	1426
Barbacenia_spectabilis	7	4694403	4692306	17844	5250	0.009489	0.000480	5029	858
Barbacenia_tomentosa	7	619710	619403	5946	3613	0.013195	0.002966	3480	1298
Barbacenia_vandellii_1	7	948595	948156	7794	4390	0.011729	0.002695	4244	1476
Barbacenia_vandellii_2	7	254909	254796	3786	2713	0.008693	0.002533	2664	1132
Vellozia_auriculata	7	1701689	1700919	9388	4017	0.013333	0.001841	3838	251

ANEXO 6. Arquivo de parâmetros utilizado no alinhamento que apresentou melhores.

------- ipyrad params file (v.0.7.28)---
barbaceniaoutgvaurifinal2 ## [0] [assembly_name]: Assembly name. Used to name output directories for assembly steps
/Users/plentbio/Desktop/victorss/ipyradprocessing/finalruns ## [1] [project_dir]: Project dir (made in curdir if not present)
/Users/plentbio/Desktop/victorss/floragenex/UO_C291_1.gz ## [2] [raw_fastq_path]: Location of raw non-demultiplexed fastq files
/Users/plentbio/Desktop/victorss/floragenex/multibarcodes.txt ## [3] [barcodes_path]: Location of barcodes file
[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
rad ## [7] [datatype]: Datatype (see docs): rad, gbs, ddrad, etc.
[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)
6 ## [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
0.82 ## [14] [clust_threshold]: Clustering threshold for de novo assembly
0 ## [15] [max_barcode_mismatch]: Max number of allowable mismatches in barcodes
0 ## [16] [filter_adapters]: Filter for adapters/primers (1 or 2=stricter)
50 ## [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)
8, 8 ## [20] [max_Hs_consens]: Max Hs (heterozygotes) in consensus (R1, R2)
4 ## [21] [min_samples_locus]: Min # samples per locus for output
24, 24 ## [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)
p, s, v	## [27] [output_formats]: Output formats (see docs)
	## [28] [pop_assign_file]: Path to population assignment file=

ANEXO 7. Relatório referente ao modelo utilizado na busca de máxima verosimilhança relatados pelo software raxml.

Tabela 2. Resultados sobre a inferência de variação dos pares de base.

	Alinhamento 1	Alinhamento 2
rate A <-> C:	1.032106	1.030770
rate A <-> G:	3.181215	3.157670
rate A <-> T:	1.054431	1.037727
rate C <-> G:	1.310563	1.299362
rate C <-> T:	3.257607	3.217691
rate G <-> T:	1.000000	1.000000
freq pi(A):	0.303557	0.303732
freq pi(C):	0.198192	0.198157
freq pi(G):	0.201680	0.201213
freq pi(T):	0.296571	0.296898

Fonte: Autor