

Julia Faillace Thiesen

**DETECÇÃO DE MARCADORES MOLECULARES  
PLASTIDIAIS PARA ESTUDOS FILOGENÉTICOS E cpSSRs  
ESPECÍFICOS PARA ORQUÍDEAS SAPATINHOS-DE-VÊNUS  
(CYPRIPEIDIOIDEAE, ORCHIDACEAE)**

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação do Curso de Ciências Biológicas da Universidade Federal de Santa Catarina em cumprimento a requisito parcial para obtenção do grau de bacharel em Ciências Biológicas. Este trabalho teve orientação da Professora Doutora Marisa Santos e co-orientação dos doutores Freek T. Bakker, Barbara Gravendeel e Rutger Voos (WUR e NBC, Países Baixos).

Florianópolis  
2015

Ficha de identificação da obra elaborada pelo autor  
através do Programa de Geração Automática da Biblioteca Universitária da  
UFSC.

A ficha de identificação é elaborada pelo próprio autor  
Maiores informações em:  
<http://portalbu.ufsc.br/ficha>

Julia Faillace Thiesen

**DETECÇÃO DE MARCADORES MOLECULARES  
PLASTIDIAIS PARA ESTUDOS FILOGENÉTICOS E  
cpSSRs ESPECÍFICOS PARA ORQUÍDEAS  
SAPATINHOS-DE-VÊNUS (CYPRIPEDIOIDEAE,  
ORCHIDACEAE)**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de ‘Bacharel em Ciências Biológicas’, e aprovado em sua forma final pelo Programa Curso de Ciências Biológicas.

Florianópolis, 19 de novembro de 2015.

---

Prof.<sup>a</sup> Dr.<sup>a</sup> Maria Risoleta Freire Marques  
**Coordenadora do Curso**

**Banca Examinadora:**

---

Prof.<sup>a</sup> Dr.<sup>a</sup> Marisa Santos  
**Orientadora**  
BOT-CCB-UFSC

---

Prof.<sup>a</sup> Dr.<sup>a</sup> Mayara Krasinski Caddah  
**Membro titular**  
BOT-CCB-UFSC

---

Ms. Lilian Machado  
**Membro titular**  
CCA - UFSC

---

Eng. Agr. Vinicius Vilperte  
**Suplente**  
CCA - UFSC



## AGRADECIMENTOS

Gostaria de agradecer imensamente aos meus supervisores Freek T. Bakker e Barbara Gravendeel, que foram muito receptivos e me deram a oportunidade de realizar um projeto de pesquisa com orquídeas durante meu programa de intercâmbio nos Países Baixos. O profundo interesse, experiência e entusiasmo em relação ao assunto e também a paciência e confiança ao me orientar, contribuíram consideravelmente à minha experiência acadêmica. Eu também gostaria de agradecer a supervisão de Rutger Voos.

Gostaria também de expressar minha admiração e gratidão à minha orientadora Marisa Santos, que acompanha minha formação acadêmica desde o princípio. Agradeço a confiança e orientação durante estes anos, bem como o paciente trabalho de revisão da redação deste estudo no pouco tempo que lhe coube, suas correções e incentivos.

Um agradecimento especial para Mathijs Nieuwenhuis, Rens Holmer e Youri Lammers, pela paciência e pelo suporte em bioinformática. Também devo agradecer a Izan Shairul Ramlee, Mark Whitten e Norris Williams por fornecer dados de sequenciamento de orquídeas, que enriqueceram este trabalho.

Ao grupo de pesquisa em biosistemática da Universidade de Wageningen e ao Centro de Pesquisa em Biodiversidade *Naturalis*, o meu muito obrigada. Agradeço ao projeto Ciências Sem Fronteiras (CNPq), por viabilizar a experiência do intercâmbio, que foi muito importante no meu processo de amadurecimento pessoal e acadêmico.

Agradeço a meu pai por seu apoio, e também a paciência, compreensão e incentivos de toda minha família e amigos.



*“ O que importa não é a perfeição com a qual possamos executar o que há de ser realizado, mas que aquilo que há de vir à vida, por mais imperfeito que venha, que seja feito em algum momento, para que haja um início!”*

*Rudolf Steiner*

*“ Nicht auf die Vollkommenheit in der wir ausführen Können dasjenige, was gewollt werden muss, kommt es an, sondern darauf, dass das, was hier ins Leben treten muss, auch wenn es noch so unvollkommen ins Leben tritt, einmal getan wird, dass ein Anfang gemacht wird!”*

*Rudolf Steiner*





## RESUMO

Apesar dos recentes avanços em genética de populações e filogenética de orquídeas, a demanda por novos marcadores moleculares ainda é alta. Orquídeas sapatinhos-de-vênus (Cypripedioideae) são exemplo de um grupo em que os atuais marcadores moleculares são insuficientes para distinguir espécies que divergiram recentemente e especificar identidade taxonômica e origem de material ilegalmente comercializado. No presente estudo, teve-se como objetivo obter o máximo de sequências do genoma do cloroplasto (cpDNA) de espécimes de orquídeas, possibilitando a detecção de novos marcadores moleculares para a subfamília. Foram processados dados de sequenciamento genômico de três orquídeas sapatinhos-de-vênus: *Cypripedium calceolus*, *Phragmipedium longifolium* e *Paphiopedilum barbatum*. As amostras foram previamente sequenciadas por autores distintos, utilizando-se de três técnicas diferentes, Illumina HiSeq 2000 e Roche 454 *Next Generation Sequencing* (NGS) e Sanger *First Generation Sequencing* (FGS). Material vegetal fresco e de herbário foram utilizados. A montagem do genoma do cloroplasto (cpDNA) foi realizada com o uso de diferentes *pipelines* de bioinformática, destacando-se o *software Iterative Organelle Genome Assembly* (IOGA). Sequências do cpDNA de material vegetal fresco e de herbário foram obtidas e análise comparativa foi realizada. Um total de 29 marcadores filogenéticos e 54 marcadores de genética de populações foram detectados. Os métodos de sequenciamento genômico e de bioinformática utilizados neste trabalho fornecem novo escopo para rápida detecção de novos marcadores moleculares a baixos custos, tanto para material fresco como de herbário.

**Palavras-chave:** *Paphiopedilum*, plastoma, *Phragmipedium*, *Cypripedium*, DNA histórico.



## ABSTRACT

Despite major advances in orchid phylogenetics and population genetics over the last decade, the demand for new genetic markers is still high. The horticulturally popular slipper orchids are a prime example of a group for which the currently available phylogenetic and population genetic markers are insufficient to resolve clades of recently diverged species and specify the taxonomic identity and origin of illegally traded material. In this study we compare previously sequenced chloroplast genomes of a total of five slipper orchid individuals. The specimens are distributed over three genera, *Cypripedium calceolus*, *Phragmipedium longifolium* and *Paphiopedilum barbatum*. The samples were sequenced by different authors and using three different techniques, Illumina HiSeq2000 and Roche 454 Next Generation sequencing and Sanger First Generation sequencing. Chloroplast genome (cpDNA) assembly was performed using a combination of bioinformatics pipelines. Comparative analysis of the retrieved sequences was performed and a total of 29 putative new phylogenetic and 54 potential population genetic markers were retrieved. The combined sequencing and assembly methods employed provide new scope for the fast detection of new markers at low costs from both fresh and museum material.

**Key words:** *Paphiopedilum*, plastome, *Phragmipedium*, *Cypripedium*, historical DNA.



## LISTA DE FIGURAS

<b>Figura 1</b> - Cladograma dos grupos principais de Orchidaceae.....	25
<b>Figura 2</b> - Cladograma de Orchidaceae.. .....	26
<b>Figura 3</b> - Estrutura quadripartida do genoma do cloroplasto.....	29
<b>Figura 4</b> - Modelo de estrutura do genoma de cloroplasto .....	30
<b>Figura 5</b> - Distribuição dos cinco gêneros de orquídeas sapatinhos-de-vênus..	33
<b>Figura 6</b> - Filogenia de orquídeas sapatinhos-de-vênus.....	34
<b>Figura 7</b> - Etapas de sequenciamento Illumina HiSeq (1).....	39
<b>Figura 8</b> - Etapas de sequenciamento IlluminaHiSeq (2) .....	40
<b>Figura 9</b> - Preparação de biblioteca genômica para 454 Roche. ....	42
<b>Figura 10</b> - Ligação de fragmentos de DNA em micro-esferas. ....	44
<b>Figura 11</b> - Roche 454 <i>pyrosequencing</i> .. .....	45
<b>Figura 12</b> - Analogia a métodos NGS.....	46
<b>Figura 13</b> - Tipos de <i>reads</i> .....	48
<b>Figura 14</b> - Exemplo do formato FASTA .....	49
<b>Figura 15</b> - Uma leitura em um arquivo FASTQ. ....	50
<b>Figura 16</b> – FASTQC (1).....	51
<b>Figura 17</b> – FASTQC (2).....	52
<b>Figura 18</b> - Relatório FASTQC.. .....	53
<b>Figura 19</b> - Relatório PRINSEQ (1). ....	54
<b>Figura 20</b> - Relatório PRINSEQ (2) .....	55
<b>Figura 21</b> - Representação de <i>readpool</i> .....	56
<b>Figura 22</b> - <i>de novo assembly</i> . ....	58
<b>Figura 23</b> - <i>Reference assembly</i> .....	60
<b>Figura 24</b> - Etapas utilizadas no <i>software</i> IOGA (BAKKER et al., 2015).....	61
<b>Figura 25</b> - Alinhamento de dois contigs. ....	62
<b>Figura 26</b> – Detecção de cpDNA <i>reads</i> .....	64
<b>Figura 27</b> - Etapas do processamento de dados NGS realizados neste trabalho. .....	72
<b>Figura 28</b> - Mapa circular do plastoma de <i>Phragmipedium longifolium</i> .....	79
<b>Figura 29</b> - a-e. Dados de cobertura de montagens do cpDNA.....	82
<b>Figura 30</b> - a-d. Espécimes disponíveis de <i>Paphiopedilum barbatum</i> (1). ....	97
<b>Figura 31</b> - e-f. Espécimes disponíveis de <i>Paphiopedilum barbatum</i> (2).....	98
<b>Figura 32</b> - i. Espécime disponível de <i>Paphiopedilum barbatum</i> (3) .....	99



## LISTA DE TABELAS

<b>Tabela 1:</b> Comparação de plataformas de sequenciamento NGS.....	36
<b>Tabela 2:</b> Comparação de diferentes atributos e comprimentos .....	47
<b>Tabela 3:</b> Detalhes das amostras .....	68
<b>Tabela 4:</b> Processamento de dados NGS: Etapa 1. ....	73
<b>Tabela 5:</b> Processamento de dados NGS: Etapa 2. ....	74
<b>Tabela 6:</b> Processamento de dados NGS. Etapas 3 .....	76
<b>Tabela 7:</b> Sequências de cpDNA montadas a partir de dados NGS. ....	81
<b>Tabela 8:</b> Marcadores filogenéticos putativos para orquídeas sapatinhos-de-vênus.....	84
<b>Tabela 9:</b> Marcadores para genética de populações de <i>P. barbatum</i> . . . . .	89
<b>Tabela 10:</b> Espécimes de <i>Paphiopedilum barbatum</i> pertencentes a coleções. . . . .	95





## LISTA DE ABREVIATURAS E SIGLAS

APG – *Angiosperm Phylogeny Group*  
BAC – *Bacterial Artificial Chromosome*  
CITES – *Convention on International Trade in Endangered Species of Wild Fauna and Flora*  
CDS – *Coding sequences*  
cp – *Chloroplast*  
cpDNA – *chloroplast DNA*  
DOGMA – *Dual Organelle Genome Annotator*  
FGS – *First Generation Sequencing*  
Gaps – *Lacunae*  
IGS – *Intergenic Spacers*  
IOGA – *Iterative Organelle Genome Assembly*  
IR – *Inverted Repeat Region*  
ITS – *Internal Transcribed Spacer*  
LSC – *Long Single Copy*  
mtDNA – *DNA mitochondrial*  
NBC – *Naturalis Biodiversity Center*  
NCBI – *National Center for Biotechnology Information*  
NOR – *Nuclear Organizer Region*  
nDNA – *DNA nuclear*  
NGS – *Next Generation Sequencing*  
OGDRAW – *Organelle Genome Draw*  
pb – *pares de base*  
PCR – *Polymerase Chain Reaction*  
SSC – *Small Single Copy*  
SSR – *Simple Sequence Repeat*  
WUR – *Wageningen University e Research Center*



## SUMÁRIO

<b>1. INTRODUÇÃO.....</b>	<b>21</b>
<b>2. OBJETIVO GERAL .....</b>	<b>23</b>
2.1. OBJETIVOS ESPECÍFICOS.....	23
<b>3. REVISÃO DE LITERATURA .....</b>	<b>24</b>
3.1. BIOSISTEMÁTICA .....	24
3.1.1. Estado da arte: filogenética de Orchidaceae.....	24
3.1.2. Genoma plastidial .....	28
3.1.3. Novos marcadores para filogenética de orquídeas.....	30
3.1.4. Cypripedioideae: inferência filogenética utilizando dados moleculares.....	31
3.2. TECNOLOGIAS DE SEQUENCIAMENTO.....	35
3.2.1. Illumina HiSeq2000 .....	38
3.2.2. Roche 454.....	41
3.3. MÉTODOS DE BIOINFORMÁTICA .....	45
3.3.1. Reads.....	47
3.3.2. <i>Genome assembly</i> : montagem do genoma .....	55
3.3.3. <i>Scaffolding</i> : agrupamento de <i>contigs</i> em <i>scaffolds</i> .....	62
3.3.4. Inferência de qualidade de montagens/ avaliação métrica .....	63
3.3.5. Acabamento do genoma .....	65
<b>4. MATERIAL E MÉTODOS .....</b>	<b>67</b>
4.1. MATERIAL VEGETAL .....	67
4.1.1. Extração de DNA .....	67
4.1.2. PCRs e preparação de biblioteca genômica .....	67
4.2. ANÁLISE DE DADOS .....	69
4.1.1. Sequenciamento Sanger.....	69
4.1.2. Dados de Sequenciamento de Nova Geração .....	69
<b>5. RESULTADOS E DISCUSSÃO .....</b>	<b>78</b>
5.1. SEQUÊNCIA COMPLETA DO PLASTOMA DE <i>PHRAGMIPEDIUM</i> <i>LONGIFOLIUM</i> .....	78

5.2. SEQUÊNCIAS PLASTIDIAIS OBTIDAS .....	80
5.3. MARCADORES FILOGENÉTICOS PUTATIVOS PARA ORQUÍDEAS SAPATINHOS-DE-VÊNUS .....	<b>83</b>
5.4. SSRs PUTATIVOS PARA <i>PAPHIOPEDILUM BARBATUM</i> .....	87
5.5. MÉRITOS SOBRE A UTILIZAÇÃO DE DIFERENTES ESTRATÉGIAS DE SEQUENCIAMENTO E MONTAGEM DE CPDNA.....	92
5.6. VALOR ACRESCENTADO PELA UTILIZAÇÃO DE ESPÉCIMES DE HERBÁRIO .....	94
<b>6. CONCLUSÕES.....</b>	<b>100</b>
<b>REFERÊNCIAS .....</b>	<b>101</b>

## 1. INTRODUÇÃO

A inferência filogenética utilizando marcadores moleculares revolucionou o entendimento de evolução de Angiospermas (CHASE et al., 1993; SOLTIS, SOLTIS e CHASE, 1999; SOLTIS, SOLTIS e ZANIS, 2002; SOLTIS et al. 2011). No entanto, a geração de marcadores moleculares para plantas era dispendiosa e demorada, pois muitas espécies possuem genoma extenso e poliploidia, requerendo a criação de bibliotecas BAC (*Bacterial Artificial Chromosome*). Durante as duas últimas décadas, por consequência, as atenções se dirigiram somente a um limitado número de marcadores de DNA nuclear e plastidial (por exemplo, os espaçadores transcritos internos do DNA ribossômico nuclear e regiões gênicas plastidiais *matK* e *rbcL*). Os progressos em técnicas de sequenciamento de nova geração (*Next Generation Sequencing* - NGS) na última década permitiram a realização de estudos evolutivos e estudos comparativos em larga escala, que antes nem se poderia imaginar (METZKER, 2010). Estas técnicas elevaram a obtenção de marcadores filogenéticos para plantas a um estágio superior, produzindo informações genômicas acuradas, rápidas e a baixos custos (METZKER, 2010). Comparações de genomas plastidiais inteiros ou sequências concatenadas de marcadores têm sido utilizadas com objetivo de aumentar a resolução filogenética em diversos níveis em Angiospermas: desde estudos a nível de ordem, como dentro de Angiospermas (CHANG et al., 2006; SHAWN et al., 2007; SOLTIS et al., 2011), até identificação de espécies e estudos genômicos populacionais (DOORDUIN et al., 2011; JHENG et al., 2012; PAN et al., 2012; YANG et al., 2013).

O uso de DNA histórico em genômica também se tornou possível com métodos NGS e genomas completos de espécimes de coleções de plantas e animais são sequenciados cada vez com maior sucesso (STAATS et al., 2013). Organismos preservados desidratados têm altos níveis de degradação do DNA, prejudicando Reações de Polimerização em Cadeia (*Polymerase Chain Reaction* - PCR), que são cruciais para o sequenciamento Sanger. As técnicas NGS, no entanto, não dependem de grandes fragmentos de DNA, mas de fragmentos curtos de 100-300pb (STAATS et al., 2013). Dados de sequência confiáveis podem ser obtidos a partir de espécimes de herbário (STAATS et al., 2011), proporcionando grande oportunidade de explorar DNA histórico em um contexto filogenético. Acervos de herbário do mundo todo podem agora prover

material para estudos moleculares e até mesmo espécies já extintas podem ser incluídas em estudos evolutivos.

Orchidaceae é uma das maiores famílias entre as plantas vasculares, que compreende 800 gêneros de plantas herbáceas terrestres ou epífitas distribuídas por todos os continentes, exceto na Antártida (FAY e CHASE, 2009). As relações filogenéticas dentro da família Orchidaceae constituem um tema importante nos estudos atuais de biologia evolutiva, pois a família representa uma ampla gama de estratégias de vida, morfologia floral e vegetativa e síndromes de polinização (FAY e CHASE, 2009). Como a família inclui muitas plantas ornamentais e medicinais, a intensa coleta ilegal reduziu drasticamente as populações naturais ao longo dos últimos 200 anos (SUBEDI et al., 2011; GHORBANI et al., 2014). A necessidade de novos marcadores para estudos filogenéticos e de populações para as orquídeas é, portanto, grande e severamente influenciada pelo fato de não haverem muitos genomas de orquídeas inteiramente sequenciados. Estes são grandes e muitas vezes poliplóides, o que dificulta o sequenciamento de genomas inteiros em orquídeas. Estudos de sistemática trouxeram respostas valiosas para classificação de orquídeas (CAMERON et al., 1999; CHASE et al., 1994; FAY e CHASE, 2009). No entanto, apesar do advento da NGS, marcadores genéticos informativos ainda não foram detectados para diversas subfamílias. Para orquídeas sapatinhos-de-vênus (Cypripedioideae), novos marcadores filogenéticos e genéticos populacionais são especialmente cruciais. Ainda não há estudos que utilizem técnicas NGS e apenas marcadores clássicos têm sido usados para esta subfamília até o momento (ver COX et al., 1997; CHOCHAI et al., 2012; GUO et al., 2012).

O presente estudo comparou sequências plastidiais de cinco indivíduos de Cypripedioideae, compreendendo três gêneros distintos (*Cypripedium*, *Phragmipedium* e *Paphiopedilum*), com objetivo de detectar novos marcadores filogenéticos e de populações para a subfamília.

Este trabalho foi desenvolvido durante o programa Ciências Sem Fronteiras (CNPq) nos Países Baixos, em período de estágio (seis meses) no Grupo de Pesquisa em Biosistemática da Universidade de Wageningen (*Wageningen University & Research Center* - WUR), sob orientação do Dr. Freek T. Bakker (WUR) e co-orientação da Dra. Barbara Gravendeel e do Dr. Rutger Voos (*Naturalis Biodiversity Center*- NBC).

## 2. OBJETIVO GERAL

O presente trabalho teve como objetivo geral estudar técnicas de montagem de genoma de cloroplasto e realizar análise comparativa de sequências plastidiais de espécimes de herbário e espécimes frescos da subfamília Cypripedioideae (Orchidaceae), visando a detecção de marcadores moleculares informativos para estudos filogenéticos e de genética de populações.

### 2.1. OBJETIVOS ESPECÍFICOS

- Recuperar o máximo de sequências plastidiais de quatro indivíduos de sapatinhos-de-vênus a partir de dados de sequência gerados anteriormente.
- Comparar as sequências do plastoma de espécimes de herbário de *Paphiopedilum barbatum* com sequências de espécimes frescos da mesma espécie.
- Produzir um alinhamento incluindo todas as sequências montadas e a sequência completa do plastoma de *Phragmipedium longifolium*, para inferir rápida evolução em regiões codificantes e não-codificantes do cpDNA.
- Produzir um alinhamento incluindo somente indivíduos de *Paphiopedilum barbatum*, para detectar e caracterizar marcadores plastidiais que possuam variação suficiente para distinguir três espécimes de *Paphiopedilum barbatum*.

### 3. REVISÃO DE LITERATURA

#### 3.1. BIOSISTEMÁTICA

##### 3.1.1. Estado da arte: filogenética de Orchidaceae

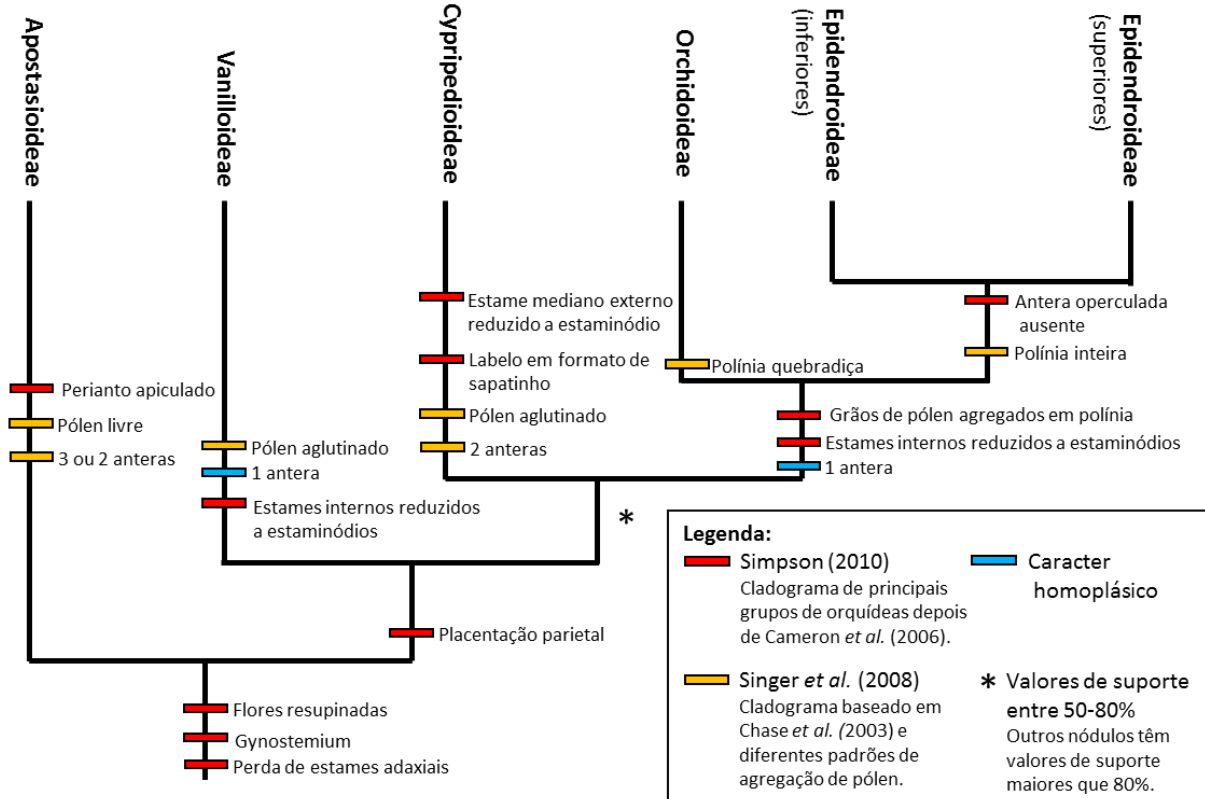
Orchidaceae é uma das maiores famílias entre as plantas vasculares, compreendendo 880 gêneros de plantas herbáceas terrestres ou epífitas distribuídas em todos os continentes, exceto Antártica (STEVENS, 2001 em diante; FAY e CHASE, 2009). Além da sua beleza e valor ornamental, a diversidade de estratégias de vida, morfologia floral e vegetativa, bem como síndromes de polinização, as fazem um objeto valioso para estudos filogenéticos. No entanto, por causa de seu tamanho, o estudo desta família em um quadro evolutivo é um desafio (FAY e CHASE, 2009): atualmente existem 27.800 espécies descritas (STEVENS, 2001 em diante). As relações filogenéticas dentro da família Orchidaceae estão se tornando bem compreendidas (SINGER et al., 2008), sendo que sinapomorfias morfológicas e moleculares suportam a divisão do grupo em cinco subfamílias: Apostasioideae, Vanilloideae, Cypridioideae, Orchidoideae e Epidendroideae (Figuras 1 e 2).

As cinco subfamílias de Orchidaceae têm sido classificadas com base tanto em dados moleculares como dados morfológicos, entretanto, ainda existem incertezas sobre a topologia da sua árvore filogenética (por exemplo, a colocação de Cypridioideae). Embora Linneus tenha sido o primeiro a publicar nomes binomiais para orquídeas, descrevendo 69 espécies em sete gêneros na primeira edição de *Species Plantarum* (1762-3), a sistemática de Orchidaceae começou com a classificação proposta por Swartz em 1805 (STERN, 2014). Swartz elaborou uma chave de identificação para 28 gêneros e dividiu dois grandes grupos baseando-se na observação de que algumas orquídeas possuíam duas anteras (diandras) e outras apenas uma (monandras). O grupo de orquídeas monandras foi dividido em três grupos baseado na posição da antera, grupos estes que hoje são conhecidos por Vanilloideae, Orchidoideae e Epidendroideae. Entretanto, a maioria das hipóteses recentes (CHASE et al., 2003; CAMERON, 2006; SINGER et al., 2008) posicionam Cypridioideae como clado irmão aos clados Orchidoideae e Epidendroideae, o que implica em considerar que a perda de um estame tenha ocorrido duas vezes (ver Figura 1 e Figura 2).

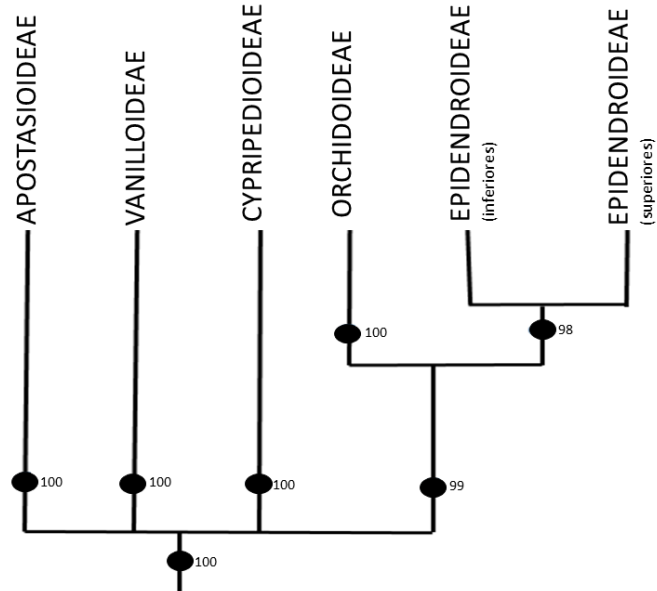
Por muitos anos relações filogenéticas de orquídeas eram inferidas utilizando-se, quase exclusivamente, características florais



**Figura 1** - Cladograma dos grupos principais de Orchidaceae. Padrões de topologia de Simpson (2010), Cameron (2006), Singer et al. (2008), Chase et al. (2003) e dados do Grupo de Filogenia de Angiospermas, ou *Angiosperm Phylogeny Group* (STEVENS, 2001 em diante) foram integrados no cladograma.



**Figura 2-** Cladograma de Orchidaceae. Árvore filogenética de consenso estrito para os genes *rbcL* e *matk* concatenados (FREUDENSTEIN et al., 2004). Valores demonstram suporte ‘*Jackknife*’. Somente clados com valores de suporte maiores que 50% são mantidos na topologia.



Fonte: FREUDENSTEIN et al. (2004).

(CAMERON, 2007). Embora dados morfológicos tenham produzido uma hierarquia apropriada em Orchidaceae, a ocorrência de homoplasia conduz em muitos casos a equívocos. Um exemplo é o estudo de Chase et al. (2009), no qual foi substituído um conjunto de espécies em *Gomesa*, previamente consideradas *Oncidium*. Além disso, filogenias de Orchidaceae fundamentadas somente em morfologia limitaram o número de caracteres concatenados e proporcionaram menos suporte para os clados (SINGER et al., 2008).

Dados moleculares têm revolucionado estudos filogenéticos em Orchidaceae desde o início da década de 1990 (FAY e CHASE, 2009). Os primeiros estudos filogenéticos baseados em caracteres moleculares (CHASE et al., 1994; CAMERON et al., 1999; FREUDENSTEIN et al., 2004) produziram topologias congruentes em comparação às hipóteses anteriores baseadas em caracteres morfológicos (DRESSLER, 1993).

Cameron et al. (1999) explorou as evidências do gene plastidial *rbcL*, usando parcimônia, e produziu um alinhamento contendo 171 táxons distribuídos em todas as cinco subfamílias, totalizando 1320 caracteres (485 filogeneticamente informativos). Comparando com Dressler (1993), a topologia gerada foi predominantemente congruente, embora a ‘elevação’ de orquídeas Vaniloides para *status* de subfamília consistiu em uma alteração importante em relação à topologia aceita até então. Freudenstein et al. (2004) expandiram o estudo de Cameron et al. (1999) concatenando sequências de *matK* às sequências de *rbcL*: o alinhamento continha 173 táxons, totalizando 2958 caracteres (1180 informativos). Os valores de suporte apresentados neste estudo são mais pronunciados do que o estudo anterior (Figura 1). Existem hoje numerosos estudos que aumentaram a resolução desta topologia e possibilitaram um quadro filogenético cada vez mais detalhado para orquídeas (FAY e CHASE, 2009).

Marcadores alternativos aos marcadores clássicos (como *rbcL* e *matK*) estão sendo sugeridos para a reconstrução filogenética em Orchidaceae (PAN et al., 2012; JHENG et al., 2012; YANG et al., 2013). O estudo de Freudenstein et al. (2004) trouxe maior suporte para os clados Epidendroideae e Orchidoideae em relação ao estudo de Cameron et al. (1999). No entanto, a topologia dos três clados basais, ou seja, dos três clados que divergiram primeiro, ainda não pôde ser elucidada, havendo uma politomia na topologia da árvore filogenética (Figura 2). Um panorama histórico de estudos filogenéticos dentro de Orchidaceae e a primeira década de sequenciamento de DNA é resumido em Cameron et al. (2007). O autor enfatiza atualizações importantes sobre a topologia da

árvore e destaca como a comparação de caracteres moleculares está transformando e aprimorando a taxonomia do grupo. *Genera Orchidacearum* (PRIDGEON et al., 1999, 2001, 2003, 2005, 2009, 2014) é apresentada como uma classificação temporária, pois novos estudos estão constantemente sendo incorporados. Uma revisão mais recente que aborda evolução no campo de filogenética de orquídeas é proporcionado por Fay e Chase (2009). O progresso dos métodos de sequenciamento têm transformado as dimensões em biologia comparativa, permitindo alinhamentos robustos de genomas completos e exploração de melhores marcadores em diferentes níveis filogenéticos.

### 3.1.2. Genoma plastidial

Técnicas de NGS têm sido utilizadas para sequenciar genomas inteiros de cloroplastos ('plastomas', ou 'cpDNA') em plantas vasculares. Sequências inteiras de genomas de cloroplastos da subfamília Epidendroideae podem ser encontrados na plataforma online *GenBank* NCBI ([www.genbank.org](http://www.genbank.org)): *Phalaenopsis aphrodite* (CHANG et al., 2006), *Phalaenopsis equestris* (JHENG et al., 2012), *Erycina pusilla* (PAN et al., 2012), *Cymbidium spp.* (YANG et al., 2013) e *Oncidium Gower Ramsey* (WU et al., 2010). A sequência do plastoma da orquídea heterotrófica *Rhizanthella gardneri* (DELANNOY et al., 2011), subfamília Orchidoideae, também pode ser encontrada na plataforma *GenBank*. Estes autores têm explorado o genoma plastidial de orquídeas, verificando que, além de sua estrutura global conservada, há regiões denominada 'hotspots' (como diferentes padrões de perda e truncagem nas subunidades da família *ndh*, padrões de *indels* em pseudogenes *ndh*, variações nas junções dos IR, SNPs e certos espaçadores gênicos). Estas regiões evoluem mais rapidamente e são muito informativas para estudos filogenéticos e de genética de populações (PAN et al., 2012).

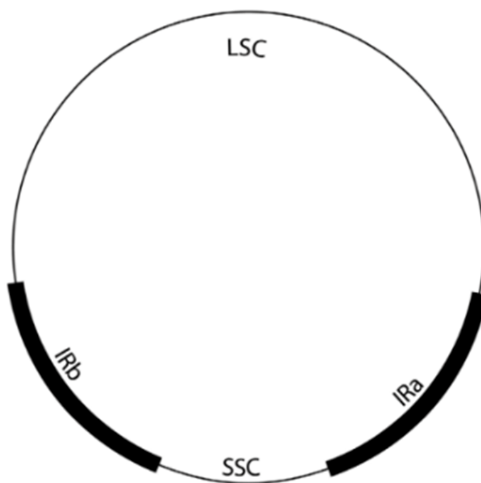
O DNA de cloroplasto tornou-se amplamente utilizado em estudos filogenéticos, principalmente devido a (1) terem várias cópias, (2) não haver riscos de paralogia e (3) herança uniparental (JUDD et al., 2002). Apesar de sua estrutura global conservada, são encontradas no genoma do cloroplasto de orquídeas regiões informativas não-codificantes (PCN) que evoluem rapidamente (conforme Jheng et al. 2012, Pan et al. 2012, Yang et al. 2013).

Segundo a teoria da endossimbiose, proposta por Lynn Margulis (1967), o cloroplasto é o produto da fagocitose de uma bactéria fotossintetizante por uma célula eucaritote. Um dos argumentos que

suporta esta teoria é a alta similaridade entre a organela e certas bactérias. O genoma do cloroplasto possui uma estrutura quadripartida muito conservada, composta por duas regiões repetidas repetidas (*Inverted Repeats* - IRs), alternadas com uma região curta de cópia simples (*Small Single Copy region* - SSC) e uma região longa de cópia simples (*Long Single Copy region* - LSC) (Figura 3). O genoma do cloroplasto está representado de forma circular (Figura 3).

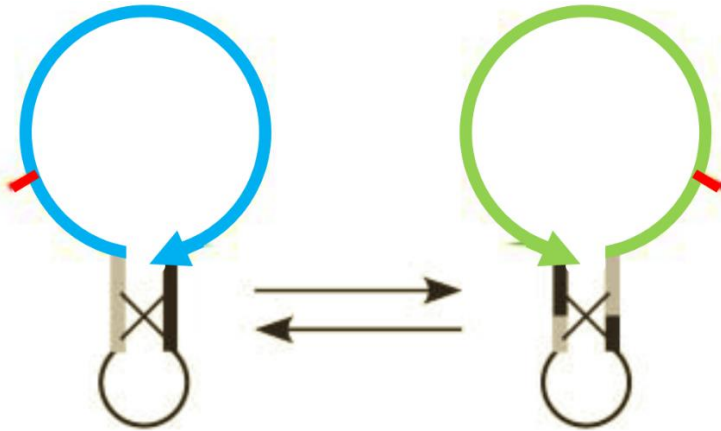
Apesar da forma circular ser amplamente utilizada, estudos apontam que esta não é a forma mais incidente (BENDICH, 2004). Por muito tempo acreditou-se que o genoma do cloroplasto era encontrado predominantemente em estrutura circular, porém estudos apontam que, assim como os cromossomos mitocondriais, o cromossomo do cloroplasto é encontrado de variadas formas e tamanhos, mostrando maior complexidade (BENDICH, 2004). A estrutura quadripartida de cloroplastos é explicada no estudo de Bendich (2004) como sendo uma molécula linear em formato de *flip flop*, tendo padrões de recombinação e reparo especiais para as IRs (Figura 4). O autor denominou este formato de *flip flop*, pois lembra um formato de chinelo.

**Figura 3** - Estrutura quadripartida do genoma do cloroplasto. Simplificado de Pan et al. (2012). Legenda: LSC – *Long Single Copy*; SSC – *Small Single Copy*; IRa e b – *Inverted Repeats*.



Fonte: Pan et al. (2012).

**Figura 4** - Modelo de estrutura do genoma de cloroplasto proposto por Bendich (2004). Este esquema é baseado inteiramente em análises estruturais de genomas de cloroplasto, e propõe que o genoma esteja organizado de maneira linear em formato de *flip flop*. Esta Figura mostra recombinações intramoleculares entre duas formas isoméricas do genoma de cloroplasto.



Fonte: Bendich et al. (2004).

### 3.1.3. Novos marcadores para filogenética de orquídeas

Jheng et al. (2012) realizaram análise comparativa entre três sequências completas do genoma do cloroplasto de orquídeas pertencentes à subfamília Epidendroideae. Esta análise comparativa revelou que, apesar do genoma plastidial ter uma estrutura geral conservada, há altos níveis de variação que são potencialmente informativas para estudos filogenéticos e populacionais. Variação no padrão de perda e truncagem de subunidades de genes e pseudogenes da família *ndh*, variação no padrão de *indels* (inserções e deleções) em regiões de subunidades *ndh*, variação nas bordas das regiões invertidas (IRs), polimorfismo de nucleotídeo único (*Single Nucleotide Polymorphism* - SNP), demonstram níveis diferentes de variações e são exemplos de *hotspots*.

Yang et al. (2013) realizaram análise comparativa de genomas plastidiais completos de oito indivíduos do gênero *Cymbidium* (Epidendroideae), detectando diferenças significativas no nível de variação entre regiões do genoma plastidial. A região SSC apresentou o maior nível de variação, com 3,5% de potenciais caracteres informativos,

seguida por regiões espaçadoras e *introns* (3,2%). As regiões LSC e IRs também apresentaram nível moderado de variação genética, com 2,7% e 0,9% de potenciais caracteres informativos, respectivamente (YANG et al., 2013). Os autores também realizaram análise filogenética para testar diferentes regiões plastidiais, comparando 10 orquídeas de diferentes grupos e sugerem a utilização de 11 sequências intergênicas plastidiais para futuros estudos filogenéticos em Orchidaceae: *cemA-petA*, *clpP-psbB*, *ndhF-rpl32*, *petA-psbJ*, *psbA-trnK*, *trnL-ccsA*, *rpl32-trnL*, *trnE-trnT*, *trnK-rps16*, *trnP-psaJ*, e *trnT-trnL*, juntamente com o marcador intergênico já bastante utilizado *trnH-psbA*. Ao explorar o genoma do cloroplasto de *Phalaenopsis aphrodite*, Chang et al. (2006) sugeriram, como marcadores promissores pra estudos filogenéticos a níveis intrafamiliares, os *introns* dos genes *rps16* e *trnK*. As sequências intergênicas *trnF-ndhJ* e *trnH-psbA* demonstraram ser altamente informativas para esclarecer relações filogenéticas dentro da subtribo Oncidiinae (Epidendroideae) (PAN et al., 2012).

### 3.1.4. Cyripedioideae: inferência filogenética utilizando dados moleculares

A subfamília Cyripedioideae contém cinco gêneros: *Selenipedium* Rchb.f., *Cyripedium* L., *Mexipedium* V.A.Albert e M.W.Chase, *Phragmipedium* Rolfe, e *Paphiopedilum* Pfitzer. *Paphiopedilum* é o maior gênero de Cyripedioideae, com 62 espécies listadas no Website do Grupo de Filogenia de Angiospermas (*Angiosperm Phylogeny Group Website*) (STEVENS, 2001 em diante), seguido por *Cyripedium*, com 46 espécies (FATIHAH, FAY e MAXTED, 2011). Limites genéricos baseados em morfologia foram definidos baseados em uma combinação de tipo de folha, venação (arranjo das folhas jovens num corpo foliar antes de abrirem), número de lóculos e placentação. Cox et al. (1997) delimitaram limites genéricos usando o marcador molecular amplamente utilizado em Angiospermas, o espaçador transcrito interno (ITS) da região do DNA ribossômico, juntamente com dados não-moleculares, utilizando parcimônia (pesagem sucessiva). Embora Cox et al. (1997) tenham fornecido uma topologia geral entre os gêneros para Cyripedioideae, o marcador usado (ITS rDNA) não foi filogeneticamente informativo além do nível de gênero, possivelmente porque a região evoluiu 'lentamente'. Além disso, a utilização de ITS rDNA como um marcador em estudos filogenéticos pode fornecer artefatos, uma vez que podem estar presentes em diversos 'ribotipos' na

região organizadora nuclear (NOR) (ÁLVAREZ e WENDEL, 2003). Portanto, a comparação de genes parálogos, que tiveram histórias evolutivas distintas, pode ocorrer. A ocorrência de homoplasia, alterações em bases compensatórias, e problemas no alinhamento devido à acumulação de *indels* e erros de sequenciamento são consequências destas comparações errôneas (ÁLVAREZ e WENDEL, 2003).

Genes nucleares e plastidiais têm sido utilizados para a realização de análises filogenéticas para a subfamília Cyprapedioideae e a topologia da mesma tornou-se relativamente bem compreendida (GUO et al., 2012). No entanto, ainda restam muitas lacunas e contradições em alguns aspectos da topologia. Por exemplo, baseando-se em marcadores morfológicos (PFITZER, 1903) e moleculares, como o gene plastidial *rbcL* (ALBERT, 1994) e o espaçador nuclear ITS rDNA (COX et al., 1997), o gênero *Selenipedium* foi colocado como grupo irmão de outras orquídeas sapatinhos-de-vênus. Por outro lado, estudos que utilizaram genes nucleares de baixa cópia *xdh* (CAMERON, 2006), genes plastidiais *atpB*, *matK* e *rbcL* (FREUDENSTEIN et al., 2004) e um conjunto de sequências plastidiais concatenadas (*matK*, *rbcL*, *rpoc1*, *rpoc2*, *ycf1*, *ycf2*) e *exons* nucleares de baixa cópia (ACO e *Leafy*) (GUO et al., 2012) suportaram a hipótese do gênero *Cyprapedium* como clado irmão de orquídeas sapatinhos-de-vênus.

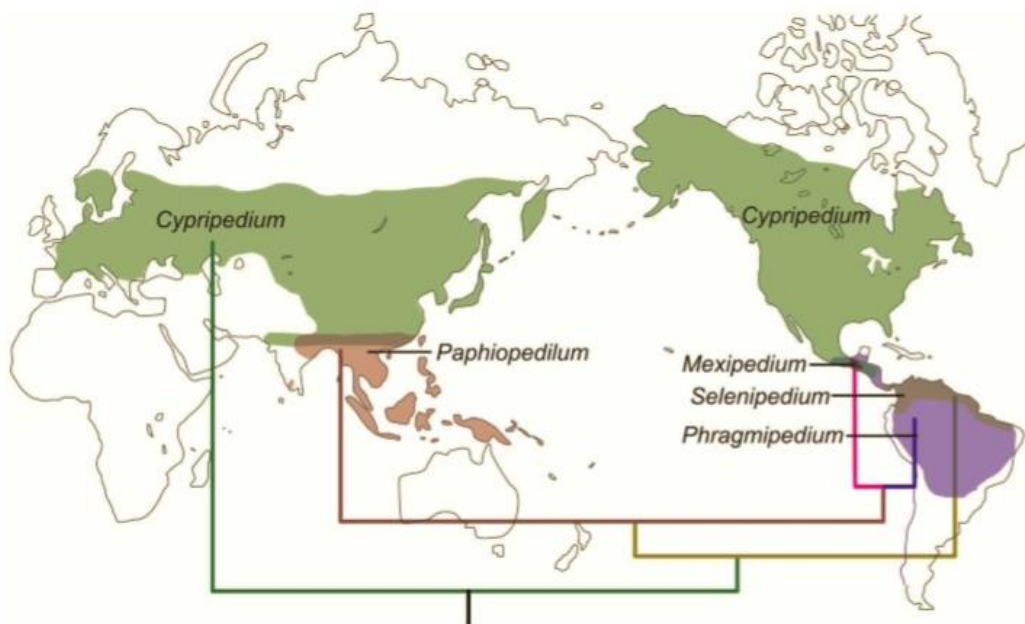
Guo et al. (2012) relatam um amplo e interessante estudo biogeográfico e filogenético de Cyprapedioideae, mostrando um padrão de distribuição disjunto e atribuindo possíveis datas de diversificação dos seus grupos. Segundo os autores, o gênero *Cyprapedium* foi o primeiro a divergir do grupo e, atualmente, tem ampla distribuição nas zonas subtropicais e temperadas no hemisfério norte. O segundo grupo a divergir foi o gênero *Selenipedium*, que atualmente é endêmico à América do Sul. *Meximedium*, *Phragmipedium* e *Paphiopedilum* estão unidos e compartilham venação conduplicada (arranjo conduplicado das folhas jovens, antes de abrirem) e distribuírem-se nas regiões dos trópicos. *Meximedium* e *Phragmipedium* são grupos mais próximos, distribuindo-se nas regiões dos neotrópicos e formam clado irmão a *Paphiopedilum*, que distribuiu-se pela Ásia tropical. De acordo com estimativas de relógios moleculares, o gênero *Selenipedium* originou-se no Paleoceno e o ancestral comum mais recente de orquídeas sapatinhos-de-vênus com folhas conduplicadas poderia ser datado para o Eoceno. Reconstrução da área ancestral indica que a vicariância é responsável pela distribuição disjunta de *Meximedium*, *Phragmipedium* e *Paphiopedilum*, nas regiões



paleotropicals e neotropicals. A Figura 5 mostra a distribuição contemporânea dos cinco gêneros através dos continentes.

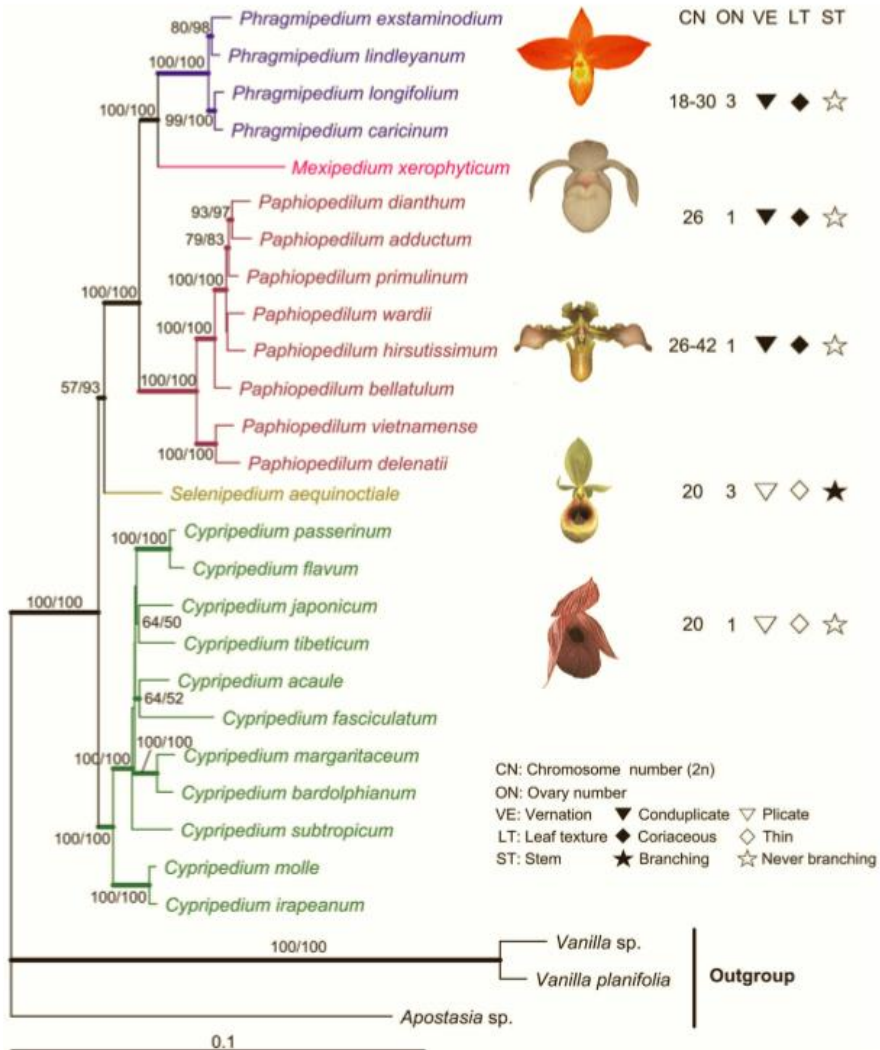
Guo et al. (2012) usaram métodos de inferência Bayesiana e de verossimilhança (*Bayesian analysis* e *Maximum Likelihood*) para analisar relações filogenéticas e biogeográficas entre 31 espécies de *Cypripedioideae*. Os métodos Bayesianos e *Maximum likelihood* são exemplos de métodos que utilizam dados discretos para calcular árvores filogenéticas (BALDAUF, 2003). Os marcadores moleculares utilizados foram sequências plastidiais concatenadas (*matK*, *rbcL*, *rpoc1*, *rpoc2*, *ycf1*, *ycf2*) e *exons* nucleares de baixa cópia (*ACO* e *Leafy*) (GUO et al., 2012). As Figuras 5 e 6 mostram a topologia obtida pelo autor.

**Figura 5** - Distribuição dos cinco gêneros de orquídeas sapatinhos-de-vênus pelos continentes (GUO et al., 2012).



Fonte: Guo et al. (2012).

**Figura 6** - Filogenia de orquídeas sapatinhos-de-vênus. Análises bayesiana, máxima parcimônia e verossimilhança, utilizando genes plastidiais e nucleares (GUO et al., 2012). Valores de probabilidade posterior ( $\geq 0.90$ ) são representadas por linhas grifadas e valores de bootstrap ( $\geq 50\%$ ) são mostrados acima dos ramos.



Fonte: Guo et al. (2012).

O gênero *Paphiopedilum* Pfitzer é um dos gêneros mais populares e raros de orquídeas vendidos e exibidos atualmente (LI et al., 2010), sendo o maior e mais 'derivado' grupo em Cyripedioideae. A última monografia do gênero foi realizada por Cribb (1997). O gênero compreende cerca de 77 espécies, as quais podem ser encontradas no sul da Índia, Nepal, Butão, nordeste da Índia, sul da China, sudeste da Ásia, Arquipélago de Malay, Filipinas, Nova Guiné e Ilhas Salomão (LI et al., 2010). O gênero é fascinante não só devido a seu valor ornamental e beleza, mas também pela curiosa característica de suas 'células-guarda' serem aclorofiladas, uma situação anormal em comparação com os estômatos de outras plantas (WILLIAMS, GRIVET e ZEIGER, 1983).

Populações selvagens estão sob crescente pressão de destruição do hábitat e coleta clandestina (LI et al., 2010) e 76 espécies de *Paphiopedilum* estão listados na Convenção sobre o Comércio Internacional de Espécies Ameaçadas de Fauna e Flora Selvagens (*Convention on International Trade in Endangered Species of Wild Fauna and Flora - CITES*). *Paphiopedilum barbatum*, como muitas outras espécies do gênero, está na lista de espécies de orquídeas ameaçadas de extinção (*Orchid Checklist of Endangered Species*). Neste caso, a possibilidade de se trabalhar com dados moleculares de espécies de herbário se torna ainda mais valiosa.

Apesar do surgimento de novas técnicas de sequenciamento de DNA, faltam estudos que se utilizem das mesmas para elucidar relações filogenéticas em Cyripedioideae. A nível de genética de populações, há, no entanto, dois estudos que fazem uso de técnicas NGS para detecção e caracterização de microssatélites para *Cypripedium calceolus* (MINASIEWICZ e ZNANIECKA, 2014) e *Cypripedium kentuckiense* (PANDEY e SHARMA, 2012). Estudos populacionais em Cyripedioideae, realizados com técnicas de isolamento de microssatélites usando Sequenciamento de Primeira Geração (*First Generation Sequencing – FGS*), incluem estudos com *Paphiopedilum rothschildianum* (RODRIGUES e KUMAR, 2009), *Paphiopedilum concolor* (LI et al., 2010) e *Cypripedium calceolus* (FAY et al., 2009). Não há, porém, estudos de genética de populações para *Paphiopedilum barbatum*.

### 3.2. TECNOLOGIAS DE SEQUENCIAMENTO

A ordem das bases nucleotídicas (adenina, timina, guanina e citosina) em um genoma pode hoje ser decifrada através de diferentes

técnicas de sequenciamento do genoma. Decorreram quinze anos entre a descoberta da dupla hélice do DNA por Watson e Crick (1953), Franklin e Gosling (1953) e o primeiro fragmento de DNA sequenciado (1968) (HUTCHISON III, 2007). As técnicas modernas de sequenciamento de DNA começaram com o desenvolvimento do método químico de Maxam e Gilbert (1977) e o método de terminação em cadeia de Sanger, Nicklen e Coulson (1977). A primeira sequência completa de DNA publicada foi do genoma de um bacteriófago com 48,5 kb (SANGER et al., 1982). O método Sanger apresentaram vantagens sobre o método químico, principalmente devido ao fato de que este último utiliza um tratamento com grande quantidade de produtos químicos tóxicos e radioisótopos. Por esta razão, de 1977 até o final de 1990, foram utilizadas quase que exclusivamente técnicas de sequenciamento Sanger (SCHUSTER, 2008). Apesar de grandes conquistas, incluindo sequenciamento completo do genoma humano, o método Sanger ainda é muito dispendioso, demorado e caro. As limitações dessa técnica geraram a necessidade de tecnologias novas e melhoradas para sequenciar grandes números de genomas humanos (METZKER, 2010). A elevada procura de novos métodos estimulou a criação de novas formas de tecnologias para sequenciamento e estas revolucionaram muitos campos da biologia, como biologia evolutiva, medicina diagnóstica e biotecnologia.

O método Sanger é considerado uma tecnologia de "primeira geração" e novos métodos são referidos como 'Sequenciamento de Nova Geração' (*Next Generation Sequencing* - NGS) (METZKER, 2010). NGS também é referido com Sequenciamento de Segunda geração (*Second Generation Sequencing*), pois constantemente este campo tem se ampliado e novas técnicas são desenvolvidas. As técnicas NGS constituem várias estratégias que se fundamentam em uma combinação de (1) preparação de modelos (ou *templates*), (2) sequenciamento, (3) criação de imagens e (4) métodos de montagem e alinhamento do genoma (METZKER, 2010). Diferentes plataformas de sequenciamento estão disponíveis no mercado e cada uma delas tem o seu próprio protocolo e resulta em tipos distintos de dados de sequências. A variedade de características NGS resulta em múltiplas plataformas que provavelmente vão coexistir no mercado, pois umas possuem claras vantagens sobre as outras (METZKER, 2010). A escolha da plataforma de sequenciamento baseia-se também na qualidade da pergunta sobre o genoma a ser sequenciado. Uma relação das plataformas de sequenciamento disponíveis podem ser encontrados na Tabela 1 (METZKER, 2010). A primeira plataforma NGS apareceu em 2005 com a publicação da

**Tabela 1:** Comparação de plataformas de sequenciamento NGS, de Metzker (2010). **Legenda:** \* Tamanho de leitura médio. ‡ *Fragment run.* § *Mate-pair run.* n.d. não disponível.

Plataforma	Tamanho Read	Gb/corrida	Custo da máquina (US\$)	Prós	Contras	Aplicações biológicas
Roche 454's/GS FLX Titanium	330* pb	0.45	500,000	<i>Reads</i> mais longas melhoram montagem em regiões repetitivas	Alto custo de reagentes; Altos níveis de erros em repetições de homopolímeros	<i>De novo assemblies</i> para genomas bacterianos e de insetos;
Illumina/Solexa's GA	75 ou 100 pb	18‡. 35§	540,000	Atualmente a plataforma mais utilizada	Baixa capacidade de multiplexação de amostras	Descoberta de variantes através de sequenciamento genômico total. Descoberta de genes em metagenômica
Life/APG's SOLiD 3	50 pb	30‡. 50§	595,000	Codificação de duas bases provê correção de erros inerentes	Técnica demorada	Re-sequenciamento de genomas bacterianos para descoberta de variantes
Polonator G.007	26 pb	12§	170,000	A plataforma mais barata;	<i>Reads</i> mais curtas	Resssequenciamento de genomas bacterianos para descoberta de variantes
Helicos BioSciences HeliScope	32* pb	37‡	999,000	Menos problemas com representação de adaptadores	Grandes níveis de erros comparado com outras tecnologias de terminadores reversíveis	Métodos baseados em sequencia
Pacific Biosciences	964* pb	n.d.	n.d.	Tem o maior potencial para <i>reads</i> acima de 1Kb	Maiores níveis de erros comparados outras NGS	Sequenciamento de transcriptoma;

tecnologia de sequenciamento ‘por síntese’ desenvolvida por 454 *Life Sciences*. Todas as plataformas de sequenciamento partilham passos semelhantes e uma revisão e caracterização destes passos é proporcionada no artigo de revisão de Metzker (2010). Aqui, somente as tecnologias IlluminaHiSeq e Roche454 serão descritas.

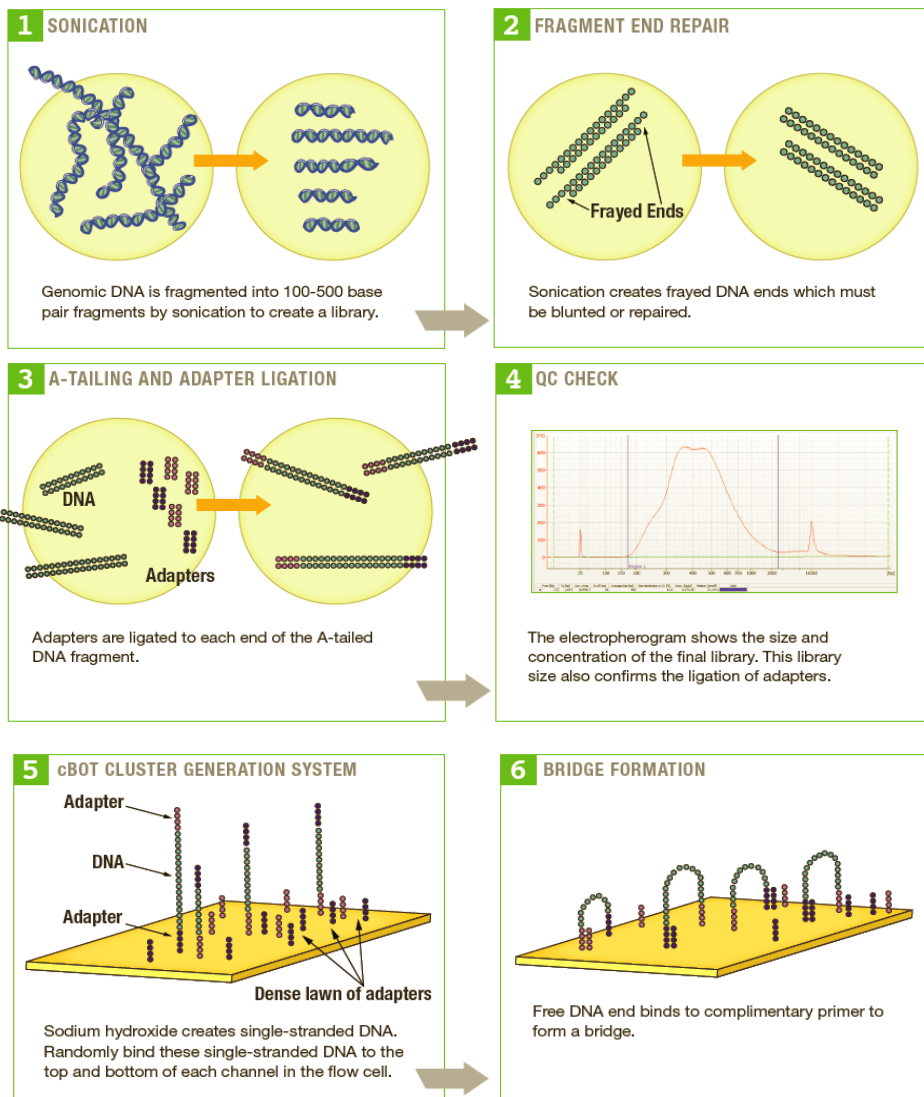
### 3.2.1. Illumina HiSeq2000

Esta descrição baseou-se em documentos disponíveis no *website* oficial da plataforma de sequenciamento Illumina ([http://www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf), acesso em 18 de junho de 2015) e em materiais disponíveis no *website* do *Department of Energy’s Joint Genome Institute* (DOEIJGI).

#### a. Preparação da biblioteca genômica

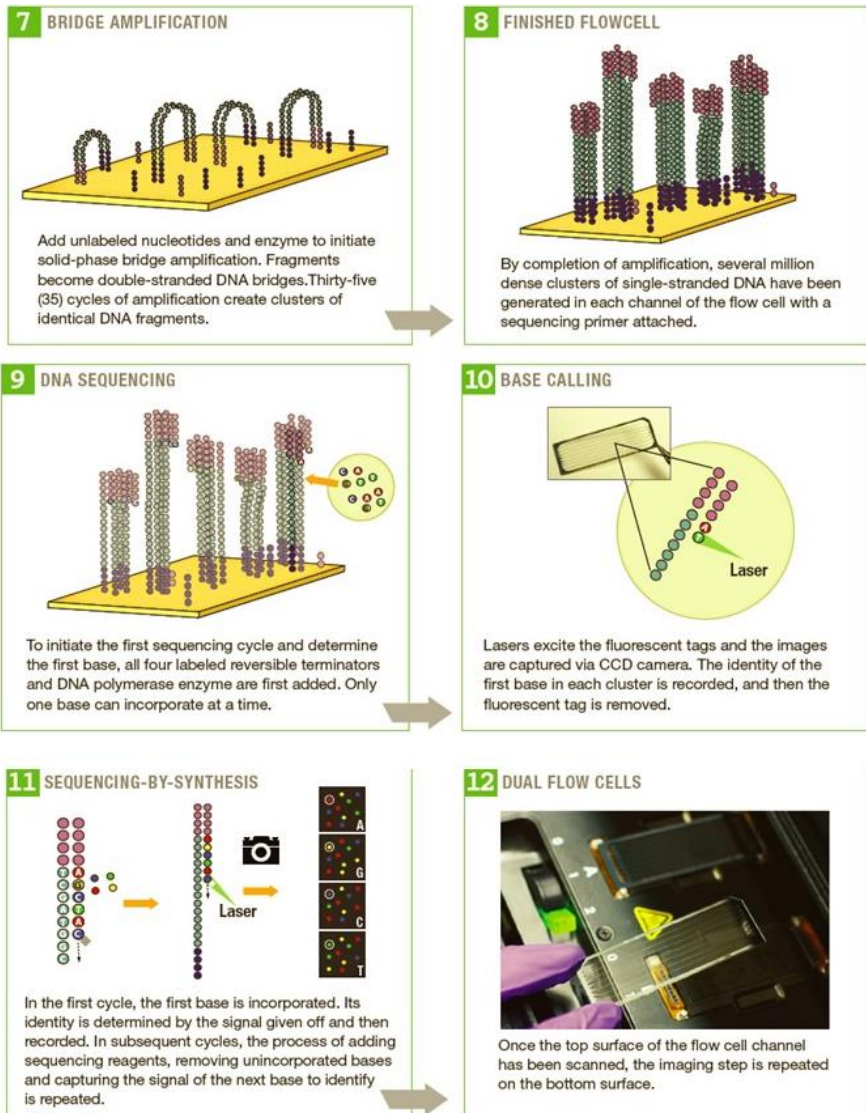
Primeiramente, o DNA é extraído e purificado do organismo em questão. Em seguida, os filamentos de DNA são fragmentados em pequenos pedaços com 100-500 pares de bases, através de sonicação (item 1 na Figura 7). A sonicação utiliza energia de ondas sonoras para desestabilizar o DNA e quebrá-lo em fragmentos. Este é um passo muito importante quando tratamos de DNA degradado, pois este já se encontra fragmentado. Para o sequenciamento Sanger, isto costumava impossibilitar ou dificultar o sequenciamento e havia necessidade de extração de grandes quantidades de DNA. Entretanto, nas técnicas de sequenciamento NGS, a fragmentação do DNA faz parte do protocolo, permitindo a otimização no sequenciamento de pequenas quantidades de DNA e gerando dados de sequência de maior qualidade. Os fragmentos de dupla hélice quando submetidos ao processo de sonicação (banho ultrassônico) sofrem desgaste das extremidades do DNA, ficando com diferentes comprimentos, o que posteriormente é reparado (item 2, Figura 7). Os fragmentos são ‘adenilados’, gerando um fragmento de DNA de cauda adenilada, e adaptadores oligos são ligados às extremidades dos fragmentos de cadeia dupla (item 3, Figura 7). A biblioteca genômica torna-se concluída para a clusterização após uma verificação de qualidade dos fragmentos, de uma seleção de tamanho de fragmentos e purificação dos mesmos (item 4, Figura 7).

**Figura 7** - Etapas de sequenciamento Illumina HiSeq. Preparação de biblioteca genômica (1-4). Passos de *clustering* (5-6).



Fonte: Dept. of Energy's Joint Genome Institute (DOE-JGI) website  
[http://openwetware.org/images/7/7a/DOE\\_JGI\\_Illumina\\_HiSeq\\_handout.pdf](http://openwetware.org/images/7/7a/DOE_JGI_Illumina_HiSeq_handout.pdf)  
 acesso em 3 de novembro, 2015)

**Figura 8** - Etapas de sequenciamento IlluminaHiSeq; passos de *Clustering* (7-8), passos de sequenciamento (9-12).



Fonte: *Dept. of Energy's Joint Genome Institute (DOE-JGI) website*  
 ([http://openwetware.org/images/7/7a/DOE\\_JGI\\_Illumina\\_HiSeq\\_handout.pdf](http://openwetware.org/images/7/7a/DOE_JGI_Illumina_HiSeq_handout.pdf)  
 acessado em 3 de novembro, 2015)



### *b. Clusterização*

Os fragmentos são isotermicamente amplificados em uma célula de fluxo e são então preparados para sequenciamento *high throughput*. Hidróxido de sódio separa os fragmentos de DNA de cadeia dupla e cria moléculas de DNA de cadeia simples. A célula de fluxo é uma lâmina de vidro com canais de compostos com dois tipos de *oligos* (também adaptadores), que estão ligados na superfície dos canais, e os fragmentos de DNA de cadeia simples são ligados de forma aleatória a estes adaptadores (item 5, Figura 7). Os fragmentos de cadeia simples são clonalmente amplificados por uma polimerase por meio de ‘amplificação por meio de ponte’ (*bridge amplification*) (item 7, Figura 8). Ao adicionar nucleotídeos e enzimas, a fase sólida da amplificação, por meio de ponte, amplifica os fragmentos de cadeia simples, sendo que durante a reação de polimerização estas pontes formam cadeias duplas, sendo separadas ao final. Trinta e cinco ciclos de amplificação criam *clusters* de filamentos de DNA idênticos. Quando a amplificação é completada, milhões de densos conjuntos de DNA de cadeia simples são gerados em cada canal da célula de fluxo, com um *primer* de sequenciamento ligado (item 8, Figura 7).

### *c. Sequenciamento*

O primeiro ciclo de sequenciamento começa após a adição das quatro bases nucleotídicas marcadas com terminadores reversíveis fluorescentes e da enzima DNA polimerase (item 9, Figura 8). Uma base é incorporada de cada vez. Lasers excitam as marcas fluorescentes e a imagem é capturada através de um sensor de uma câmera digital (item 10, Figura 8). A identidade da primeira base em cada *cluster* é registrada e então o marcador fluorescente é removido (itens 10 e 11, Figura 8). O passo de formação de imagens é repetido na superfície do fundo do canal de fluxo da célula. As seqüências nucleotídicas lidas nesta etapa são armazenados em arquivos de texto, que contém todas as leituras dos fragmentos sequenciados.

#### **3.2.2. Roche 454**

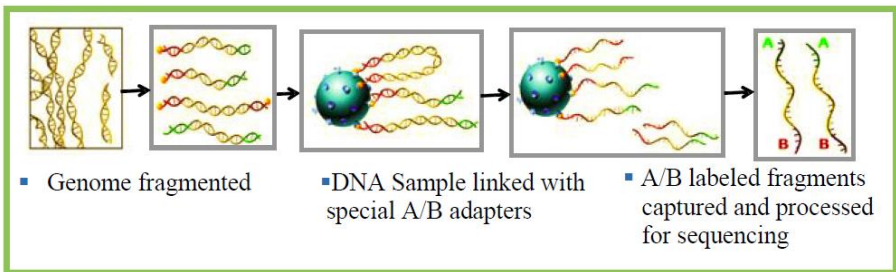
Roche 454 *pyrosequencing* é uma técnica de Sequenciamento de Nova Geração que se tornou disponível no mercado em 2005. A descrição a seguir foi baseada inteiramente em material disponível no *website* da

Roche 454 ([http://www.454.com/downloads/news-events/how-genome-sequencing-is-done\\_FINAL.pdf](http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf), acessado em 18 de junho, 2015 ).

*a. Preparação da biblioteca genômica*

Fragmentos de DNA do genoma inteiro ou genomas-alvo (cpDNA, nDNA e mtDNA) são fragmentados pela primeira vez através de sonicação ou nebulização. Fragmentos de 300-800 pares de bases são gerados e os adaptadores são ligados às extremidades dos fragmentos. Os fragmentos de DNA de cadeia dupla são separados em cadeias simples (Figura 9).

**Figura 9** - Preparação de biblioteca genômica para 454 Roche *pyrosequencing*.



Fonte: Roche 454 *website* ([http://www.454.com/downloads/news-events/how-genome-sequencing-is-done\\_FINAL.pdf](http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf), acesso em 3 de novembro de 2015).

*b. Processando amostras em micro-esferas (beads)*

Fragmentos de DNA da biblioteca são colocados em micro-esferas, através de um processo de amplificação clonal à base de emulsão (*Emulsion PCR*). Como resultado da amplificação dos fragmentos de DNA, os sinais produzidos durante o passo de sequenciamento são facilmente detectáveis. Uma micro-esfera irá conter apenas um fragmento de DNA, que fica imóvel durante as reações. Cada micro-esfera contém uma solução enzimática que realiza Reações de Polimerização em Cadeia (PCR) e multiplica o fragmento de DNA em aproximadamente 10 milhões de cópias. Quando a reação de PCR se completa, as micro-esferas são selecionadas e limpas. As micro-esferas que não produziram informações de DNA são eliminadas. As micro-esferas que possuem mais de um tipo de fragmento de DNA são facilmente filtrados durante o

processamento de sinal de sequenciamento. Este processo é exemplificado na Figura 10.

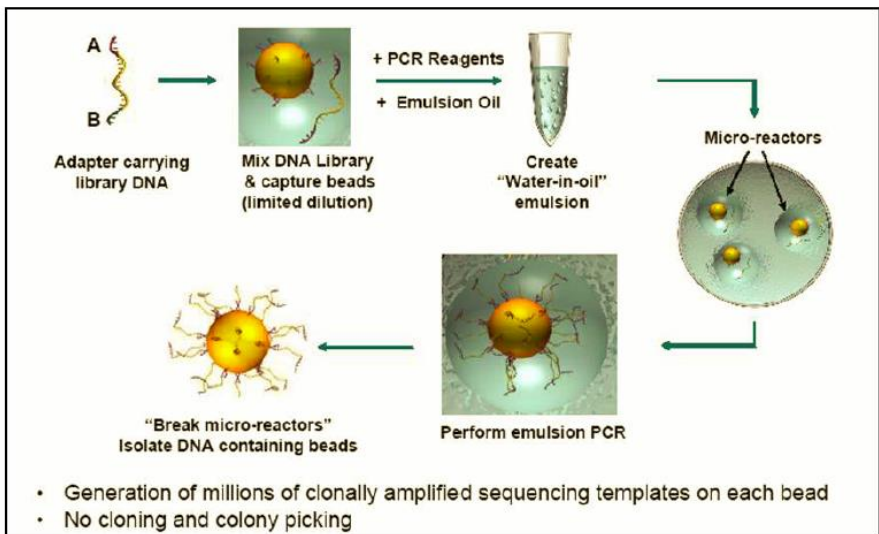
### c. Sequenciamento

O processo de sequenciamento 454 (Figuras 10 e 11) utiliza uma abordagem de sequenciamento ‘por síntese’ para gerar dados de sequência. No sequenciamento por síntese, um fragmento de DNA de cadeia simples é multiplicado através de reações de polimerização, que origina clones de fragmentos de cadeia dupla. A polimerase adiciona sequencialmente um nucleotídeo complementar à sequência do fragmento de DNA de cadeia simples. Nucleotídeos são combinados um a um à medida em que a enzima se move ao longo do fragmento de cadeia simples, estendendo a estrutura de dupla hélice. As micro-esferas de captura de DNA são colocados em uma placa, o ‘PicoTiterPlate™’ para sequenciamento. Um dos lados do PicoTiterPlate é polido e o outro lado da placa contém ‘poços’ ou ‘cavidades’ que possuem 75 picolitros de volume. Cada PicoTiterPlate compreende 1,6 milhões de poços ou reservatórios. O diâmetro dos poços é concebido de modo a que apenas uma única micro-esfera de captura se encaixa em cada cavidade. Depois que estas cavidades estão repletas de micro-esferas de captura contendo as cadeias de DNA amplificadas e fragmentadas, juntamente com muitas enzimas, a placa é colocada no instrumento sequenciador do sistema Roche 454. Este instrumento inclui um sistema de fluidos capazes de limpar a placa PicoTiterPlate com vários reagentes, incluindo nucleotídeos A, C, G e T. Os quatro nucleotídeos fluem sequencialmente em quatro lavagens sobre a placa. Quando esses nucleotídeos são incorporadas sobre as cadeias de DNA, as enzimas contidas em cada cavidade do placa convertem os produtos químicos gerados durante a incorporação de nucleotídeos em energia luminosa, numa reação químico-luminescente semelhante àquela utilizada por um vaga-lume. A incorporação de um nucleotídeo complementar de um fragmento de DNA gera um sinal luminoso, e este é detectada por uma câmera digital CCD.

A intensidade do sinal é proporcional ao número de nucleotídeos incorporados (Figura 11). Uma câmera CCD utiliza um pequeno pedaço retangular de silício, em vez de um pedaço de película, para receber a luz recebida. Esta é uma peça especial de silício chamado de ‘dispositivo de acoplamento de carga’, ou CCD (para *charge-coupled device*). A intensidade de luz gerada durante o fluxo de um único nucleotídeo varia proporcionalmente com o número consecutivo de nucleotídeos complementares no fragmento de DNA de cadeia simples a ser analisado.

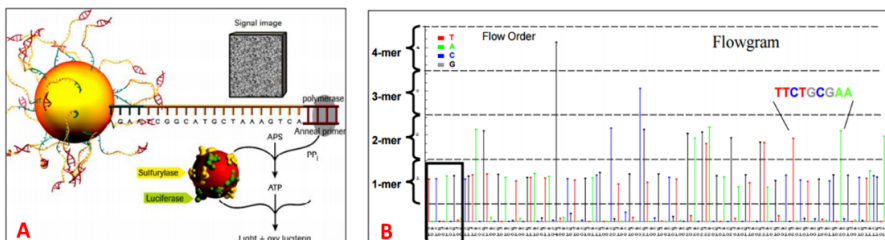
Por exemplo, se houver três adeninas consecutivas no fragmento de cadeia simples, a quantidade de luz gerada seria três vezes maior que a de um único 'A' na sequência. Os sinais criados no processo de sequenciamento são então analisados por um *software* do Sistema Sequenciador 454. É possível gerar milhões de bases sequenciadas por hora a partir de um único ciclo. A partir do sinal químico-luminescente, este programa gera um gráfico de barras de intensidade de luz chamado "Flowgram", para cada poço ou cavidade contida no PicoTiterPlate™.

**Figura 10** - Ligação de fragmentos de DNA em micro-esferas.



Fonte: *Dept. of Energy's Joint Genome Institute (DOE-JGI) website* ([http://openwetware.org/images/7/7a/DOE\\_JGI\\_Illumina\\_HiSeq\\_handout.pdf](http://openwetware.org/images/7/7a/DOE_JGI_Illumina_HiSeq_handout.pdf) acessado em 3 de novembro, 2015).

**Figura 11 - Roche 454 pyrosequencing.** A. Sequenciamento ‘por síntese’: sinal químico-luminescente é produzido na reação de síntese de nucleotídeos B. Um software analisa a intensidade de incorporação de nucleotídeos, ou seja, a intensidade do sinal luminescente. Esta intensidade informa a qualidade da leitura. O *software* analisa a intensidade do sinal luminescente, atribuindo um valor de qualidade para cada base sequenciada.



Fonte: Dept. of Energy's Joint Genome Institute (DOE-JGI) website ([http://openwetware.org/images/7/7a/DOE\\_JGI\\_Illumina\\_HiSeq\\_handout.pdf](http://openwetware.org/images/7/7a/DOE_JGI_Illumina_HiSeq_handout.pdf) acessado em 3 de novembro, 2015)

### 3.3. MÉTODOS DE BIOINFORMÁTICA

Durante o processo de sequenciamento NGS, as informações de seqüências nucleotídicas vão sendo armazenadas em arquivos de texto. No caso de dados de seqüência de DNA, os nucleotídeos estão representados pelas letras A (adenina), G (guanina), C (citosina) e T (timina). Os dados de seqüência de fragmentos são chamados de ‘*sequence reads*’, ou ‘leituras de seqüência’. No caso de sequenciamento de DNA total de um dado organismo, leituras de seqüências genômicas nucleares, mitocondriais e de cloroplastos (no caso de organismos fotossintetizantes) estão misturadas no mesmo arquivo de texto. A posição relativa das *reads* não é conhecida. O produto final de sequenciadores NGS seria como uma caixa de peças de quebra-cabeça (análogas às *reads*), que precisam de uma série de etapas de montagem para que sejam organizados de maneira correta.

Como os termos utilizados em bioinformática geralmente são utilizados no idioma Inglês, gera desafios para referir as informações. Sendo assim, os termos neste trabalho foram usados na forma original (Inglês) ou substituídos de modo a facilitar o entendimento. O formato e característica das *reads* são discutidos no item 3.3.1. A reorganização de pequenos fragmentos de leitura em seqüências mais longas e,

eventualmente, da sequência inteira do genoma é chamado de ‘*genome assembly*’, ou ‘montagem do genoma’.

Estratégias diferentes de montagens são discutidas no item 3.3.2. O produto de montagem, ou seja, a sequência produzida, também pode ser chamado de *assembly*, havendo duas etapas distintas: (1) *draft assembly*, que é um rascunho ou projeto da sequência final e (2) *final assembly*, que é a sequência final. Esta é produzida a partir da *draft assembly*, e somente é considerada a "montagem final" quando as lacunas restantes foram resolvidas (*gaps*) e a posição relativa de sequências é conhecida (após o acabamento do genoma). A Figura 12 mostra uma analogia das principais etapas de sequenciamento NGS até a montagem final (*final assembly*). Na primeira etapa, as cópias de DNA do genoma total são fragmentadas para sequenciamento. Na segunda etapa, estes fragmentos de DNA são lidos e leituras de sequências são produzidas e armazenadas. Na terceira etapa, estas leituras de sequência são reorganizadas numa sequência linear que representa o genoma do organismo sequenciado.

**Figura 12** - Analogia a métodos NGS. Pilhas de jornal representam cópias do DNA genômico total. O DNA é fragmentado em subsequências pequenas do genoma e então é sequenciado. As leituras destas subsequências são remontadas como num quebra-cabeça, até constituir uma única sequência. Modificado de Seeman (2011).



Fonte: Seeman (2011).

### 3.3.1. Reads

*Reads* são pequenas leituras do genoma sequenciados, em outras palavras, são subsequências curtas do genoma (SEEMAN, 2011). Dependendo da tecnologia de sequenciamento, *reads* têm diferentes comprimentos e atributos (SEEMAN, 2011). Todas as tecnologias NGS produzem grande número de *reads*, devido às etapas de clonagem. A plataforma de sequenciamento NGS Roche 454 produz sequências mais longas comparadas às *reads* da plataforma Illumina HighSeq, no entanto o tamanho das *reads* é mais variável. *Reads* mais extensas são mais facilmente montadas e têm uma maior capacidade de detecção de elementos repetitivos.

**Tabela 2:** Comparação de diferentes atributos e comprimentos (pb) de reads provenientes de diferentes plataformas de sequenciamento. Simplificado de Seeman (2011).

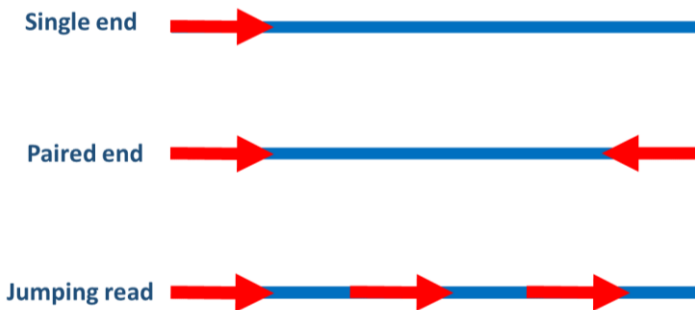
Método	bp	Atributos
Sanger	1200	Baixa quantidade de <i>reads</i> produzidas, baixa qualidade nas duas extremidades
Roche 454	700	Erros em homopolímeros
Illumina HiSeq	150	Baixa qualidade na extremidade 3'
SOLid	75	Difícil de trabalhar em <i>colour space</i>
Ion Torrent	100	Homopolímeros, altos níveis de erros
PacBio	3000	Muito alto nível de erros, pode ser iterativo

Fonte: Seeman (2011).

Tecnologias de sequenciamento NGS podem produzir três tipos de *reads*: (a) *single end read*, que poderia ser traduzido como 'leituras não pareadas'; (b) *paired end reads*, 'leituras pareadas'; (c) *strobe* ou *jumping reads*, que são *reads* de diversas orientações de um mesmo fragmento de DNA. As *reads* pareadas são produzidas quando as duas extremidades do fragmento são sequenciadas em sentidos inversos (chamamos de leitura 'em ambas direções', mas a leitura acontece a partir de diferentes extremidades do mesmo fragmento de DNA). O sequenciamento de apenas uma extremidade dos fragmento de DNA produz *reads* não pareadas. Os três tipos de *reads* são mostrados na Figura

13. A produção de *reads* pareadas (*paired end sequencing*) requer a mesma quantidade de DNA e resulta em dados de seqüências muito melhores, pois estas *reads* pareadas são montadas mais facilmente. O sequenciamento Illumina HiSeq *paired end reverse* produz dois arquivos diferentes de um mesmo ciclo de sequenciamento, um contendo as *reads* no sentido direto (*forward*) e o outro as *reads* no sentido inverso (*reverse*). Ambos devem ter aproximadamente a mesma quantidade de *reads*. Como as *reads* são como peças de um quebra-cabeça, o conjunto do genoma deveria ser um problema de simples dedução. Entretanto *reads* são relativamente muito curtas, e a realidade é que montá-las e recuperar toda a seqüência do genoma torna-se um complicado problema de inferência (SEEMAN, 2011). As *reads* podem conter erros de sequenciamento, como colocação de bases erradas pelas enzimas polimerases, colocações extras ou bases ignoradas, repetições errôneas de seqüências e outros problemas.

**Figura 13** - Tipos de *reads*.



Fonte: Modificada de Seeman (2011).

a. *Formato de arquivos de leitura*

*Reads* são salvas em um ou dois arquivos em formato de texto. No caso de dados de seqüência de DNA, os nucleotídeos estão representados pelas letras A (adenina), G (guanina), C (citosina) e T (timina). Este sistema registra tanto as seqüências nucleotídicas quanto os valores de qualidade da leitura de cada nucleotídeo, tendo sido desenvolvido para armazenar dados de seqüência produzidos com a tecnologia Sanger. No entanto, este sistema é amplamente utilizado por outras plataformas de sequenciamento, inclusive a maioria das tecnologias NGS. O sistema



Ilumina HiSeq produz *reads* com comprimento aproximado de 100 bases, os quais são normalmente fornecidos em formato de FASTQ. *Reads* podem também ser armazenadas em FASTA, que inclui somente a sequência nucleotídica e exclui qualquer qualidade associada à mesma. Ambos FASTA e FASTQ são arquivos de texto, contendo informações sobre sequências de nucleotídeos ou sequências de peptídeos.

### *Formato FASTA*

Formato FASTA é um arquivo baseado em texto contendo *reads* e identificadores (ID). É composto de duas linhas: a identificação da sequência de leitura (*read*) e sequência nucleotídica propriamente dita (Figura 14). O valores de qualidade para cada leitura de nucleotídeos é mantido em outro arquivo, o arquivo de qualidade. Estes dois arquivos podem ser fundidos em um arquivo FASTQ por certos comandos, se necessário. Se a tecnologia de sequenciamento gerar *reads* pareadas, as *reads* no sentido direto e inverso são salvas em arquivos distintos.

**Figura 14** - Exemplo do formato FASTA. A primeira a linha mostra o ID de sequência, 'SequênciaA', e as demais linhas formam uma pequena *read*.

```
>SequênciaA
AATACAATCATAATAGTTGAAAGTACCAGAGATTCCTAGAGG
CATACCATCAGAAAACTTCCTTGACCGATTGGATAGATCAA
GAACAACAGCGGTAGCAGAATAACAATCATAAAGTTGAAAGT
ACCAGAGATTCCTAGAGGCATACCATCAGAACAACTTCCTT
GACCGATTGGATAGATCAAGAAAACAGCGGTAGCAG
```

Fonte: Autor.

### *Formato FASTQ*

Arquivos FASTQ normalmente são fornecidos em formato GNUzip (.gz), que é uma compressão de FASTQ. Este formato reúne os dados de sequência e outras informações anexadas a ele, como o ID de *reads* e a qualidade de *reads* por base. Cada *read* possui 4 linhas: a primeira linha é o identificador de sequência (o nome da *read*); a segunda linha representa a sequência de nucleotídeos; a terceira linha tem um "+" e, opcionalmente, o ID de sequência novamente; e a última linha é o índice de qualidade por base. Uma leitura muito curta é mostrada na Figura 15.

**Figura 15** - Uma leitura em um arquivo FASTQ. Existem 4 linhas por leitura. O ID de sequência é a primeira linha e inicia com '@', o padrão de identificação varia com a tecnologia de sequenciamento utilizada. Esta é uma leitura Illumina HiSeq2000. A segunda linha é a sequência de nucleotídeos. Terceira linha tem sempre um '+' e pode conter ou não o ID da sequência. A quarta linha mostra a classificação de qualidade de leitura para cada nucleotídeo.

```
@HWI-ST897:243:C2RNEACXX:4:1101:11511:16883
GATGCGTCTTCTATTCTTTTCCCTGACGCAGCTGGGCCATCCT
GGACTTGAAAGATCGGAAGAGGGTCGTGTA
+
BBBFFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJEHHH
FFFBEDDDDBD
```

Fonte: autor.


#### *b. Avaliação da qualidade de reads*

Existem muitos programas *online* e também para *download* que podem avaliar a qualidade da leitura de dados NGS. Uma ferramenta online amplamente utilizada (há também uma versão para *download*) é o relatório FastQC (disponível no site <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> acessado em 3 de novembro, 2015). FastQC analisa sequência de leitura e problemas pontuais originados na preparação da biblioteca ou nas etapas de sequenciamento. Além de estatísticas básicas a seguinte informação é dada: qualidade por base na sequência, qualidade por fragmento de sequência, índices de qualidade por sequência, conteúdo de sequência por base, níveis de GC na sequência, conteúdo de base N (neutras, ou bases que não foram identificadas), distribuição de comprimentos de sequência, níveis de duplicação de sequência, repetição de sequências, conteúdo de adaptadores (sequências adaptadoras que foram adicionadas aos fragmentos de DNA durante o sequenciamento) e conteúdo de *Kmer* (subsequências de uma *read*). PRINSEQ é outro *software* muito útil e também muito similar ao FASTQC (disponível no site <http://prinseq.sourceforge.net/> acessado em 3 de novembro, 2015). PRINSEQ mostra o comprimento das sequências, qualidades de base, conteúdo GC, caudas poli-T e poli-A (contaminações por sequências adaptadoras ou erros realizados pelas polimerases), bases ambíguas e níveis de duplicações de sequência.

## FASTQC

Illumina HiSeq2000 *paired end* podem ser submetidas a uma análise FastQC, sendo fornecida uma visão panorâmica qualitativa das populações de *reads*. Podem existir diferentes populações de *reads* pois muitas vezes o sequenciador produz diferentes gradientes de qualidades de *read*, por exemplo diferentes comprimentos de *reads*. O *software* é gratuito, podendo ser obtido no site <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (acessado dia 3 de novembro de 2015), no qual há documentação consistente para compreender e interpretar os resultados de cada parâmetro. O relatório apresenta 12 ítems, alguns deles são mostrados nas Figuras 16, 17 e 18. Com esta informação, se não há uma população de *reads* com valores mais altos para posições mais terminais, poder-se-ia decidir por excluir as últimas posições das *reads*, e com isto não incluir *reads* ambíguas e de má qualidade na sua *assembly*. No entanto, pode acontecer de os dados de sequenciamento possuírem a maioria das populações de *reads* com má qualidade, porém a minoria das *reads* ainda forma uma população com alto número de *reads* de boa qualidade, não havendo necessidade de excluir as últimas posições de toda amostra.

**Figura 16** - FASTQC. Estatística básica para um arquivo contendo dados NGS produzidos por Illumina NGS. Detecção do código do sequenciador, número total de reads, número de reads de baixa qualidade, comprimento médio de reads e conteúdo de GC são fornecidos.

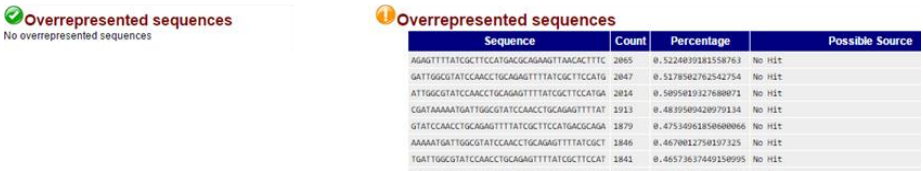


### Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

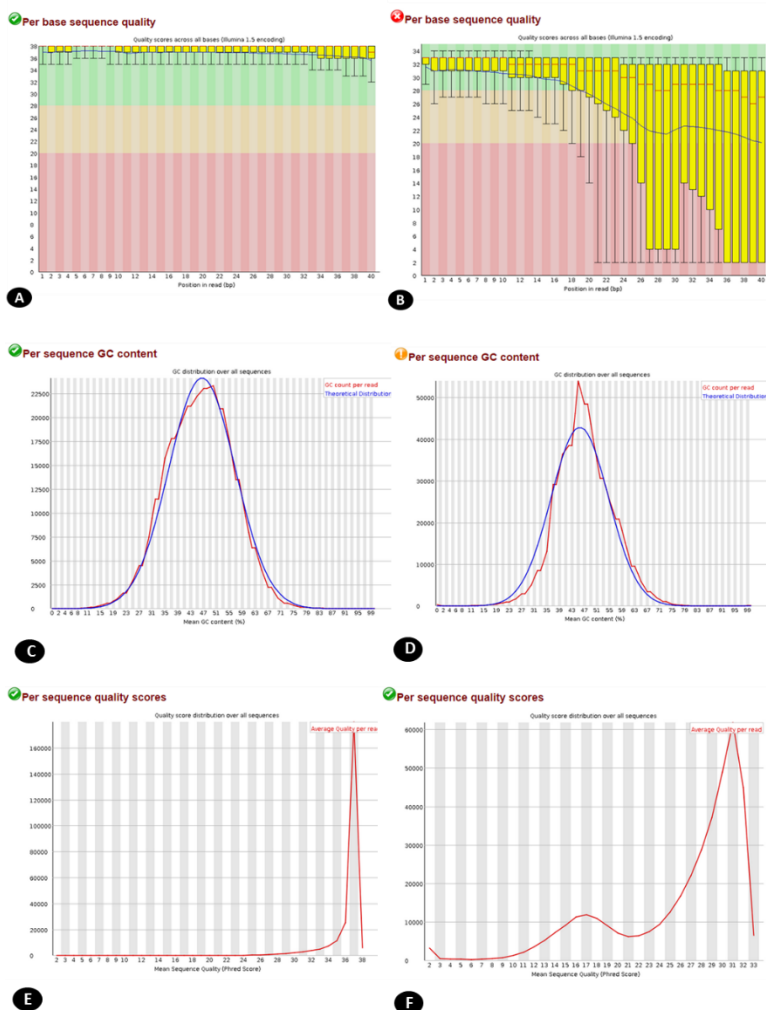
Fonte: FastQC *website* ([//www.bioinformatics.babraham.ac.uk/projects/fastqc/](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) acesso em 3 de novembro de 2015).

**Figura 17 - FASTQC.** Contaminação de amostras. Em arquivos de *reads* de má qualidade, frequentemente aparecem sequências com níveis anormais, como por exemplo com níveis exagerados, e isto pode acontecer por erros em passos do sequenciamento ou pode ter havido contaminação.



Fonte: FastqQC *website* ([//www.bioinformatics.babraham.ac.uk/projects/fastqc/](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) acesso em 3 de novembro de 2015).

**Figura 18** - Relatório FASTQC. A-B. Qualidade de sequência por base; A. amostra de boa qualidade; B. amostra de má qualidade, qualidade diminuída em posições finais dos fragmentos (pb); C-D. Conteúdo de GC por sequência; C. amostra de boa qualidade; D. amostra de má qualidade; E-F. Valores qualitativos por sequência; E. amostra de boa qualidade; F. amostra de má qualidade.



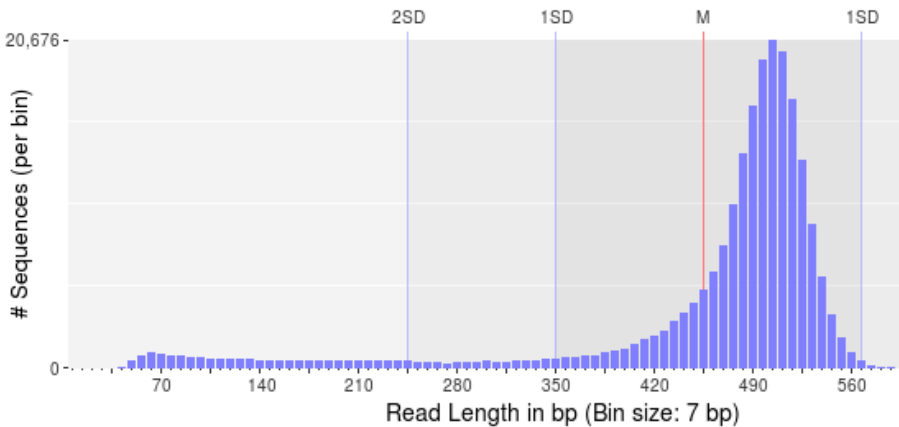
Fonte: FastqQC *website* ([//www.bioinformatics.babraham.ac.uk/projects/fastqc/](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) acesso em 3 de novembro de 2015).

*PRINSEQ*

O PRINSEQ também é um *software* de uso livre (disponível para *download* em <http://sourceforge.net/projects/prinseq/files/>, acessado dia 3 de novembro de 2015, ou também pode ser usado on-line, se o conjunto de dados não for muito pesado). Um bom manual pode ser encontrado no mesmo site (<http://prinseq.sourceforge.net/manual.html> acessado dia 3 de novembro de 2015). PRINSEQ funciona de forma semelhante ao FastQC, oferecendo uma série de análises estatísticas de FASTA e arquivos QUAL, ou mesmo arquivos FASTQ (Figuras 19 e 20).

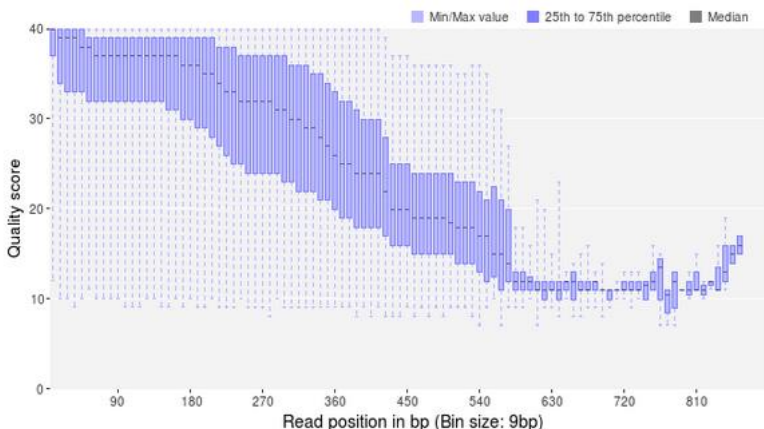
O PRINSEQ apresenta estatísticas básicas, como o número e comprimento das sequências, distribuição de comprimento (Figura 19), níveis de homopolímeros, valores de qualidade de bases por posição (Figura 20), distribuição de conteúdo GC, caudas poli-A / T, bases ambíguas, complexidade de sequência, contaminação de sequência, entre outras análises. Ainda, PRINSEQ realiza análise qualitativa e quantitativa de populações de *scaffolds* e *contigs*, conteúdo de Kmer, N50, entre outros.

**Figura 19** - Relatório PRINSEQ (1). Distribuição de tamanho de leitura (em pares de base). Amostra apresenta uma grande população entre 420 e 520 pb. Normalmente nós esperamos que a totalidade de reads que tenham aproximadamente o mesmo tamanho.



Fonte: PRINSEQ *website*, disponível em <http://prinseq.sourceforge.net/manual.html>, acessado dia 3 de novembro de 2015).

**Figura 20** – Relatório PRINSEQ (2). *Scores* de qualidade por posição de base nas reads. Dados de IlluminaHiSeq tipicamente tem baixa qualidade no final da leitura.



Fonte: PRINSEQ *website*, disponível em <http://prinseq.sourceforge.net/manual.html>, acessado dia 3 de novembro de 2015).

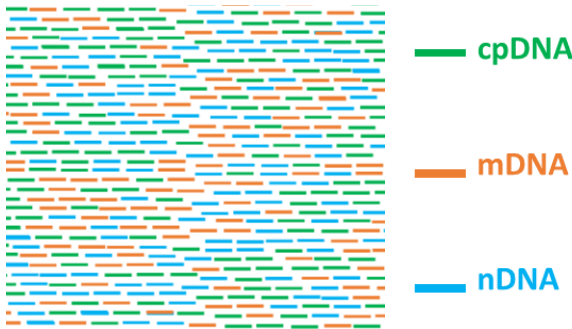
### 3.3.2. *Genome assembly: montagem do genoma*

Os dados de sequência NGS, ou seja, o conjunto total de *reads*, são salvos em arquivos FASTQ ou FASTA. Estes arquivos contêm todas as *reads* de genoma do cloroplasto, genoma mitocondrial e genoma nuclear (no caso de plantas). Este grupo de dados de sequência é o conjunto total de *reads* produzidas. Pode ser chamado de *readpool*, e é referido como 'R' em Bakker et al. (2015). *Readpool* está representado na Figura 21. Cada linha representa uma *read*. *Reads* pertencentes a genoma mitocondrial, genoma de cloroplasto e genomas nucleares são exibidos com cores diferentes (laranja, azul e verde para mtDNA, nDNA e cpDNA respectivamente). O *Readpool* contém todas as *reads* de maneira desordenada, portanto a posição relativa das mesmas é desconhecida.

A montagem do genoma (*assembly*) é normalmente realizada em duas fases: (1) *reads* são organizadas em *contigs* (que é uma sequência contígua de nucleotídeos, é o produto da combinação e sobreposição de *reads*); (2) os *contigs* são organizados em *scaffolds* através da utilização de informações de sequências pareadas (*paired end reads*) (HUNT et al., 2014). *Reads* podem ser montadas em sequências mais alongadas (*contigs*

e *scaffolds*) por programas que utilizam algoritmos (montagem *de novo*) ou podem ser alinhadas a uma sequência genômica de referência. A decisão de utilizar qualquer estratégia baseia-se na aplicação biológica a que se destina, bem como custo, esforço e considerações sobre tempo para realização (METZKER, 2010). Montagens baseadas em referência somente são eficientes se a sequência do genoma de uma espécie altamente semelhante estiver disponível. *Reads* são selecionadas de ‘R’, ou *readpool*, somente se são altamente semelhantes ao DNA de referência.

**Figura 21** - Representação de ‘R’, ou ‘*readpool*’. *Reads* do genoma nuclear, mitocondrial e do cloroplasto são mostrados em azul, laranja e verde, respectivamente.



Fonte: autor.

Uma vez que são selecionadas de ‘R’, são alinhadas com o genoma de referência. O alinhamento pode mostrar que há muitas *reads* idênticas alinhadas umas às outras na referência, o que depende principalmente da qualidade do sequenciamento e do nível de similaridade da espécie-alvo e da espécie de referência. Se as espécies tiverem alto nível de semelhança, o alinhamento mostrará muitas *reads* sobrepostas e gerando um valor de cobertura (*coverage* ou *read depth*). O valor de cobertura é um aspecto importante para se inferir qualidade em montagem de genomas. Quanto mais sobreposições de *reads* houver no alinhamento, maior o valor de cobertura e maior a confiança no posicionamento de determinada sequência.

Após a produção do alinhamento de um genoma de referência com *reads* mais conservadas de ‘R’, o consenso deste alinhamento é criado e a montagem do genoma é finalizada. O nível de semelhança deve ser



muito elevado para atingir uma boa qualidade de montagem. Assumir que a organização genômica da espécie alvo e da espécie de referência são muito similares às vezes pode gerar erros. Quando se assume que o genoma de referência é a sequência ‘correta’, somente são selecionadas *reads* contendo sequências altamente conservadas. Desta forma, rearranjos genômicos e *reads* contendo sequências que dão identidade à espécie alvo ficam excluídas. A montagem ‘*de novo*’ não exclui *reads* particulares específicas, e, portanto, tem maior poder em detectar variações. Por outro lado, as regiões repetitivas do genoma são um dos maiores problemas em *de novo*, são chamadas ‘repetições’, ou ‘*repeats*’. Repetições são segmentos de DNA que se repetem em várias regiões (por exemplo, *transposons*, inserções, duplicações, genes parálogos) (SEEMAN, 2011) e são confundidas pelos programas por serem tão similares. É impossível resolver uma repetição mais longa que o comprimento de leitura. Por isso que para detectar regiões repetitivas do DNA, que são muito informativas para genética de populações, deve-se escolher pela técnica de sequenciamento que produza sequências mais alongadas, como PCR de longo alcance (*Long Range PCR*) seguido de sequenciamento NGS.

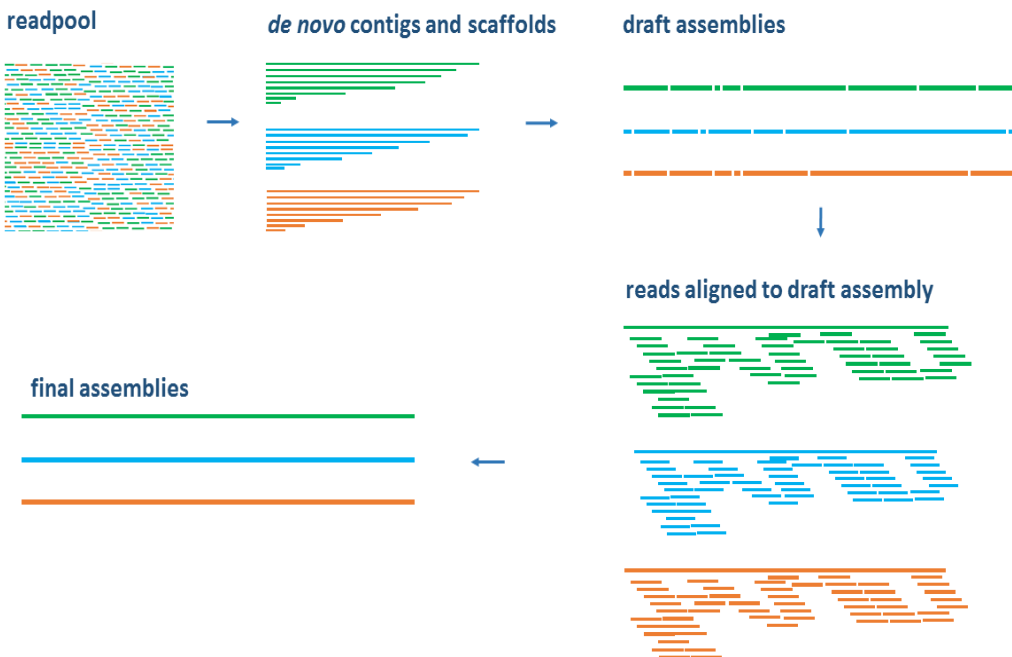
Diferentes estratégias de montagem, ou ‘*pipelines*’, são descritas abaixo (itens a, b e c). O primeiro passo útil para processar ‘R’, em qualquer *pipeline* de montagem, é filtrar *reads* de baixa qualidade e sequências de primers e adaptadores (específicas da plataforma de sequenciamento). Este processo seletivo é chamado de ‘*trimming*’, que poderia ser traduzido como ‘aparar’. Isso pode ser feito através de vários pacotes de *software*, como TRIMMOMATIC ou PRINSEQ.

#### a. *de novo assembly: montagem de novo*

A montagem ‘*de novo*’ do genoma organiza as *reads* de ‘R’ (*readpool*), utilizando apenas ‘R’ como arquivos de entrada. Há muitos programas que utilizam algoritmos para executar tal tarefa, como SOAPdenovo2 (disponível para *download* em <http://sourceforge.net/projects/soapdenovo2/files/SOAPdenovo2/>, acesso 19 de junho de 2015) e velvet (disponível para *download* <https://github.com/dzerbino/velvet/tree/master>, acesso em 19 de junho de 2015). As *reads* são combinadas entre si, formando *contigs*, que são sequências contíguas e lineares de *reads* de bases. Estes *contigs* são compostos por *reads* (*reads*) sobrepostas. Para dados de sequenciamento genômico pareados e *strobe* os *contigs* são unidos em *scaffolds*, utilizando

as informações de *reads* diretas e inversas. A Figura 22 ilustra os passos principais desta estratégia de montagem (*de novo assembly pipeline*): (1) *Reads* (*reads*) são montadas em *contigs* e *scaffolds* por programas que usam algoritmos; (2) *draft assembly* é produzida (sequência-rascunho) e as *reads* são alinhadas a esta sequência para que valores de cobertura sejam observados. Isso também pode ser feito com *scaffolds* separadamente, se não houver meios de inferir as posições relativas destes e não for possível a obtenção de uma única sequência. A montagem final, ou sequência final obtida (*final assembly*) é produzida após a inferência da posição relativa de *scaffolds* e depois de resolver lacunas (*gaps*) (por exemplo com sequenciamento complementar).

**Figura 22** - *de novo assembly*. Softwares utilizando algoritmos realizam a montagem de R (*Readpool*) em *de novo contigs* e *scaffolds*. Um alinhamento é produzido para gerar dados de cobertura, e idealmente o genoma plastidial, nuclear e mitocondrial são montados (no caso do sequenciamento de DNA total de plantas).

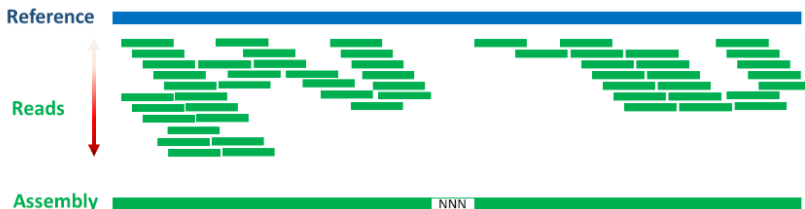


*b. Reference assembly: montagem de referência*

Na montagem de referência (Figura 23), um genoma muito semelhante é utilizado como modelos, e as *reads* são selecionadas de ‘R’ e alinhadas ao genoma de referência se há alto nível de semelhança de sequência. Se a sequência de DNA utilizado como referência é uma sequência plastidial, *reads* do genoma plastidial serão selecionadas de ‘R’. As sequências que contêm muita variação serão ignoradas, bem como inserções, deleções e rearranjos. A quantidade de variação permitida pode ser regulada por diferentes parâmetros. Pode-se incluir maior tolerância a variações, mas isto pode acarretar em erros na montagem. Os valores de cobertura variam de acordo com a quantidade de clones de *reads* (*reads*) que se sobrepõem uns aos outros em relação à referência (seta vermelha indica cobertura na Figura 23) Podemos agrupar, ou colapsar, as *reads* sobrepostas em uma sequência consensual contígua (*draft assembly* ou sequência-rascunho). Quando houver regiões sem alinhamento de *reads*, elas serão preenchidas com bases ambíguas (Ns). Estas regiões são chamadas de *gaps* ou lacunas. Estas lacunas devem ser resolvidas com sequenciamento complementar. Ao conhecer as regiões em torno de uma lacuna, *primers* podem ser projetados. O Sequenciamento Sanger ou o preparo de amostras com PCR de longo alcance (*Long Range PCR*) são muitas vezes utilizados para preencher as lacunas e detectar rearranjos no genoma.

Pacotes gratuitos de *software* que executam a montagem de referência estão disponíveis para *download*, como *Bowtie2* (disponíveis em <http://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.5/>, acessado em 19 de junho, 2015) e *Burrows-Wheeler Aligner* (BWA) (disponível em <http://sourceforge.net/projects/bio-bwa/files/>, acessado em 19 de junho de 2015).

**Figura 23 - Reference assembly.** Montagem baseada numa sequência genômica de referência. *Reads* idênticas sobrepõem-se umas às outras, reads contendo extremidades idênticas sobrepõem-se nas mesmas extremidades, cobrindo lacunas. Para gerar uma sequência conseso as informações de alinhamento são colapsadas em uma sequência linear e regiões contendo mais variações entre referência e espécie-alvo são preenchidas por bases neutras, ou “Ns”.



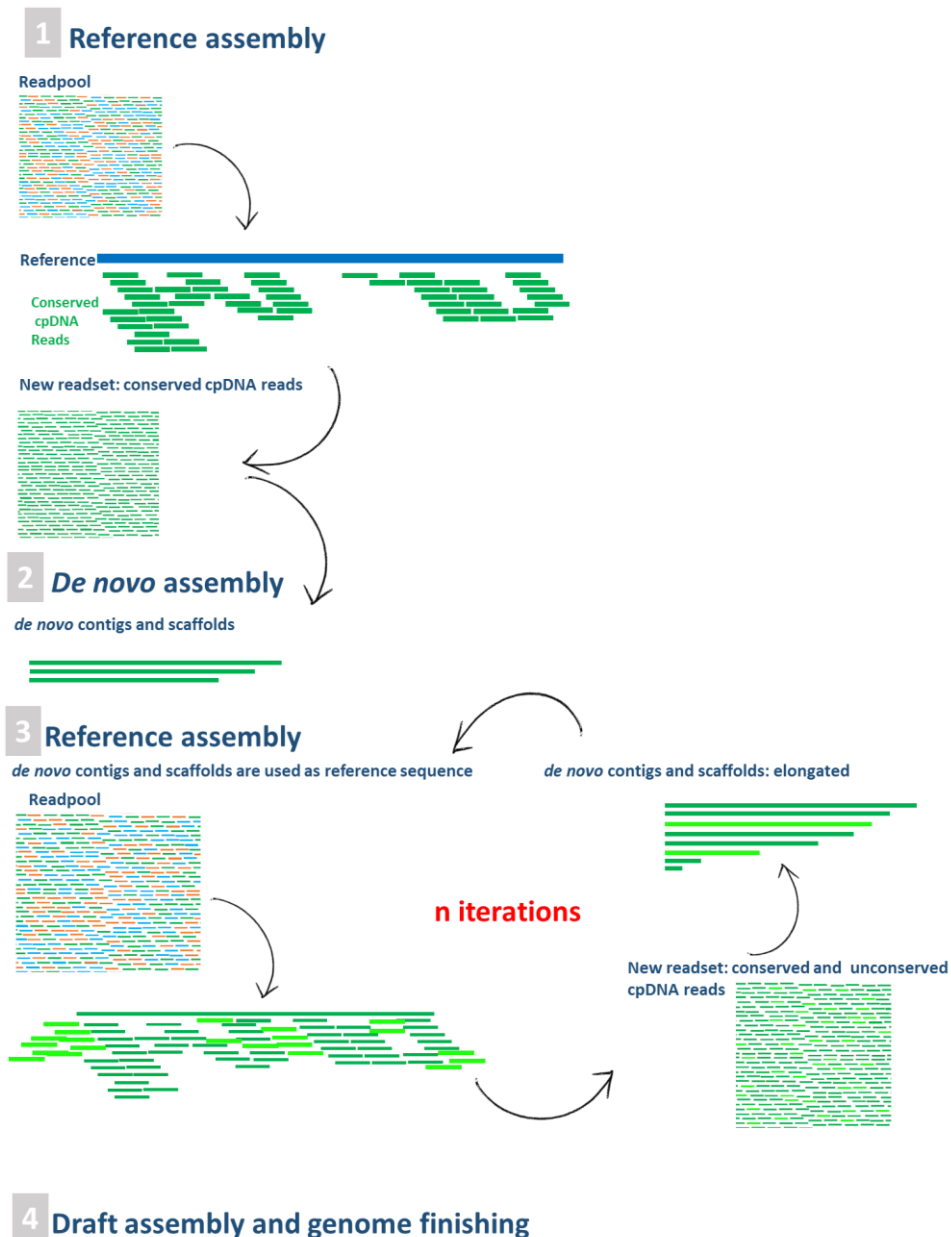
Fonte: autor.

### c. Combinações de diferentes estratégias de montagem

Ambas estratégias de montagem de genoma podem ser combinadas para gerar uma *final assembly* de maior qualidade, que mescle as vantagens de uma técnica com as vantagens da outra. Montagem baseada em um genoma de referência é mais hábil em inferir posição relativa de *reads* ou mesmo de *scaffolds* gerados *de novo*. A utilização de *contigs* e *scaffolds* gerados *de novo* possibilita a inclusão de rearranjos e *reads* que carreguem altos níveis de variação. Este método é descrito para a montagem do genoma mitocondrial em Hahn et al. (2013), e para a montagem de genoma de cloroplasto em Bakker et al. (2015).

O *software* produzido para montagem do genoma mitocondrial é chamado MITObin, sendo de livre acesso (disponível para *download* em <https://github.com/chrishah/MITObim>, acesso em 19 de junho, 2015). O *software* descrito em Bakker et al. (2015) é uma estratégia que combina *reference assembly* e *de novo assembly* de forma iterativa, e tem como nome *Iterative Organelle Genome Assembly – IOGA*, que poderia ser traduzido como Montagem Iterativa de Genoma Organelar. Este *software* também é de livre acesso (disponível para *download* em <https://github.com/holmrenser/IOGA>, acesso em 19 de outubro de 2015). MitoBIN e IOGA são softwares automatizados que utilizam uma série de programas para processar ‘R’ em um arquivo contendo candidatos a *assemblies*. Uma diferença importante entre eles é que MITObin necessita

**Figura 24** - Etapas utilizadas no *software* IOGA (BAKKER et al., 2015).



Fonte: autor

de um genoma de referência estreitamente relacionado com a amostra, enquanto que IOGA trabalha com um ‘painel de referência’. Este é um banco de dados que inclui genomas inteiros de muitas espécies de Angiospermas e serve como referência.

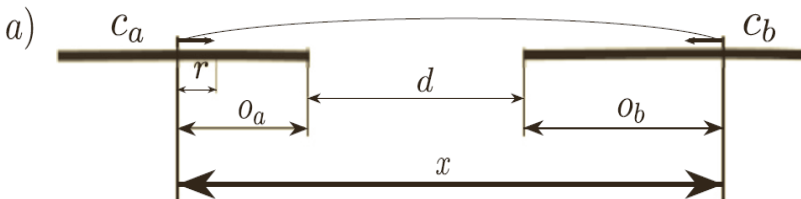
Outros softwares bastante úteis para passos intermediários de processamento dados de sequenciamento NGS são: igvtools, samtools e samstats. A iteração descrita neste dois *softwares* acima citados também poderia ser feita manualmente, mas obviamente é muito mais demorada e dispendiosa. A Figura 24 mostra os passos principais do IOGA.

### 3.3.3. *Scaffolding: agrupamento de contigs em scaffolds*

O processo de ligar e ordenar *contigs* é chamado de *scaffolding* (SAHLIN et al., 2014). O agrupamento de *contigs* fornece estimativas sobre a ordem, a orientação relativa e distância (*gaps*) entre fragmentos (SAHLIN et al., 2012). As distâncias entre *contigs* são inferidas por meio de *softwares* de *scaffolding*, ou também pode ser feito manualmente usando BLASTn e mapeando os *contigs* (se você tiver uma *assembly* não muito fragmentada) em outros genomas. As *gaps*, ou lacunas, são preenchidas com bases ambíguas quando não solucionadas.

*Contigs* gerados via *de novo* são conectados usando informações de *reads* pareadas (SAHLIN et al., 2012). Um par de leitura é definido como duas *reads* que são sequenciadas a uma distância e orientação conhecidas, onde a distância entre as *reads* é referida como ‘tamanho de inserção’ (SAHLIN et al., 2014). Alinhamento de *contigs* utilizando informação de par de leitura é ilustrada na Figura 25.

**Figura 25** - Alinhamento de dois contigs ( $C_a$  e  $C_b$ ), utilizando-se de informações de pares de sequência, extraído de Sahlin et al. (2014).



Fonte: Sahlin et al. (2014)

Existem muitos programas de agrupamento de *contigs* ('*scaffolders*'), uma avaliação dos programas atualmente utilizados pode ser encontrada em (HUNT et al., 2014). Uma nova ferramenta para agrupamento eficiente de grandes *assemblies* fragmentadas (BESST) é proposta em Sahlin et al. (2014). Duas etapas principais são comuns em programas de agrupamento: (1) otimização da ordem e da orientação de *contigs* e (2) a verificação de inconsistências de informações de *reads* pareadas. A etapa 2 envolve a remoção de *reads* com tamanho de inserção muito pequeno ou muito grande (você pode definir limites de tamanhos de inserção) e remoção de *reads* que foram mapeadas em uma posição relativa errônea (SAHLIN et al., 2012). Após estas duas etapas principais, a distância entre *contigs* é inferida, e diferentes programas de agrupamento de *contigs* usam diferentes modelos para isto. Pacotes de programas que realizam *scaffolding* variam amplamente em sua qualidade e são altamente dependentes da qualidade dos dados de leitura e da complexidade do genoma (MANDRIC e ZELIKOVSKY, 2015).

### 3.3.4. Inferência de qualidade de montagens/ avaliação métrica

O controle de qualidade ideal seria comparar os agrupamentos de montagem gerados com sequências geradas com sequenciamento Sanger. No entanto, devido aos elevados custos e tempo, outros parâmetros de controle de qualidade são usados. Algo muito útil a fazer após a montagem do genoma ser feita é mapear as *reads* (de 'R') à montagem de genoma produzido (*assembly*), produzindo um arquivo de alinhamento e sendo possível verificar os valores de cobertura. Os valores de cobertura podem ser gerados com muitos *softwares*, como SAMTOOLS e IGVTOOLS, e demonstram quantas *reads* se sobrepuseram em cada uma das posições do montagem final do genoma. Os valores de cobertura podem ser comparados com a posição das bases (em *scaffolds*) e qualidade pode ser verificada mesmo em genomas interminados. Pode-se visualizar as regiões com boa cobertura, ou com níveis exagerados de cobertura (indicando contaminação, erros de montagem ou repetição de sequências), ou baixos níveis de cobertura, ou nenhuma cobertura, indicando erros em etapas de montagem ou etapas de bioinformática ou má qualidade de extração de DNA e sequenciamento.

O tamanho total de montagem (em pares de base), o número total de *contigs* e *scaffolds* e o número total de *reads* alinhadas são boas informações para inferir a qualidade da montagem, embora não possam ser considerados separadamente de outros parâmetros. O alinhamento de





### b. *Conteúdo de GC*

O conteúdo Guanina-Citosina (teor de GC) é a proporção de bases G e C de uma sequência. O par de bases GC compartilham três pontes de hidrogênio, sendo mais estável e mais fortemente ligado do que pares de nucleotídeos AT. Consequentemente sequências de DNA com maior teor de GC são mais estáveis. Em uma sequência existem regiões de 'baixo teor GC' ou 'pobres em GC' e regiões com 'alto teor de GC'. O teor de GC médio de plastomas de angiospermas é 38-39%, e as suas regiões codificantes de rRNA e tRNA são significativamente mais ricas nos teores de GC em torno de 55% (HOLMER e NIEUWENHUIS, 2014).

### c. *Assembly Likelihood Estimator – ALE*

Assembly Likelihood Estimator, ou 'avaliador da probabilidade de acerto de montagem', é um pacote de software que avalia sistematicamente a precisão de uma *assembly* usando métodos estatísticos rigorosos (CLARK et al., 2013). Qualidade de leitura, orientação de *reads* pareadas, tamanho de inserção (para *reads* pareadas), cobertura de sequenciamento, alinhamento de *reads* e a frequência e distribuição de *k-mer* (subsequências de *reads*) estão integrados na análise ALE. ALE aponta erros sintéticos em montagens de genomas individuais e em agrupamentos metagenômicos, incluindo erros de base única, inserções/deleções, rearranjos genômicos e montagens quiméricas apresentados em metagenomas (CLARK et al., 2013). ALE combina todos esses parâmetros e gera um valor de probabilidade de acerto para uma montagem, indicando sua qualidade. Este *software* é incluído nos passos finais de IOGA (BAKKER et al., 2015).

### 3.3.5. Acabamento do genoma

Depois de conectar *de novo contigs* em *scaffolds*, o projeto de montagem (ou sequência-rascunho) está pronto. É chamado *draft assembly*. Este ainda tem *gaps* e pode estar muito fragmentado. A posição relativa de *scaffolds* ainda pode ser desconhecida. Se *scaffolding* de alta qualidade foi executado, uma boa estimativa das distâncias entre os *contigs* e *scaffolds* quantificará de maneira efetiva as lacunas do genoma, e, portanto, vai ajudar nas etapas de acabamento do mesmo (SAHLIN et al., 2012). Estimativas de má qualidade sobre distâncias de *contigs*, especialmente estimativas baixas, podem induzir erros no acabamento,

podendo também interferir na anotação gênica do genoma (SAHLIN et al., 2012). Depois que distâncias entre *gaps* são estimada, *primers* devem ser projetados em função destes e sequenciamento adicional deve ser realizado a fim de resolvê-los. O Sequenciamento Sanger é adequado para resolver pequenas lacunas (por exemplo 150pb), e outras técnicas, como *Long Range PCR* e *IonTorrent* devem resolver lacunas maiores. Depois de resolver *gaps*, o alinhamento de ‘R’ na montagem final é indicado para gerar novos valores de cobertura. Anotação de genes pode ser realizada com o DOGMA, ou Geneious.

## 4. MATERIAL E MÉTODOS

### 4.1. MATERIAL VEGETAL

O genoma total de cinco orquídeas sapatinhos-de-vênus foi sequenciado: um indivíduo de *Cypripedium calceolus*, um indivíduo de *Phragmipedium longifolium* e três indivíduos de *Paphiopedilum barbatum*. O sequenciamento foi realizado por autores distintos, utilizando-se de técnicas de NGS (Illumina HiSeq 2000 e Roche 454) e sequenciamento Sanger. Material vegetal fresco e de herbário foram utilizados. Os detalhes de cada amostra são apresentados na Tabela 3. Os dois espécimes de herbário de *Paphiopedilum barbatum* foram sequenciados como parte do projeto europeu SYNTHESYS (finalizado em janeiro de 2012, [www.synthesys.info](http://www.synthesys.info)), e como parte de um estudo mais amplo: o Grupo de Biossistemática da Universidade de Wageningen, Países Baixos, gerou dados NGS para 94 amostras frescas e históricas (BAKKER et al., 2015).

#### 4.1.1. Extração de DNA

O método CTAB de Doyle e Doyle (1987) foi utilizado para extrair DNA genômico total proveniente de amostras frescas, usando 50 mg de tecido foliar. Amostras de herbário tiveram seu DNA total extraído de 50 mg de tecido foliar desidratado em sílica, utilizando o kit comercial 'DNA Plant Mini Kit' (QIAGEN), seguindo o protocolo sugerido pela empresa. Extrações de DNA de amostras desidratadas foram realizadas numa instalação especial para DNA histórico, com objetivo de evitar contaminação com DNA moderno.

#### 4.1.2. PCRs e preparação de biblioteca genômica

##### a. Sequenciamento Sanger

Whitten et al. (não publicado) obteve a sequência completa do genoma do cloroplasto de *Phragmipedium longifolium*, gerando produtos de PCR de longo alcance (*Long Range PCR*), os quais foram sequenciados pelo método Sanger descrito em Pakendorf et al. (2006). Os autores disponibilizaram as sequências para o presente trabalho.

**Tabela 3:** Detalhes das amostras. Os cinco espécimes são provenientes de coleções diferentes e foram sequenciados com diferentes plataformas (sequenciamento não foi parte deste trabalho). **Legenda:** n.d. Dados desconhecidos. \* Valor baseado no tamanho do cpDNA de *P. longifolium*.

Espécime	Coleta	País	Coletor	n° coleção	Local de seq.	Plataforma	# reads	Tamanho médio de reads
<i>Phragmipedium longifolium</i>	2013	Equador Baños	Mark Whitten Whitten 2804	n.d.	Flórida, EUA	Sanger	1	150909 pb
<i>Cypripedium calceolus</i>	n.d.	Reino Unido	Mike Fay	Unknown	Kew, Inglaterra	Roche 454	34,048	300 a 1500 pb
<i>Paphiopedilum barbatum</i>	2013	n.d.	Anônimo	15097287	BGI, China	IlluminaHiSeq	15,097,287	~100 pb
<i>Paphiopedilum barbatum</i>	1968 (herbário)	Malásia, Pennang Hill	Anônimo 030	L0717340	<i>National High-throughput Sequencing Centre,</i> Dinamarca	IlluminaHiSeq	12,104,266	~98 pb
<i>Paphiopedilum barbatum</i>	1970 (herbário)	Malásia, Terengganu	C. Davidson 1248	L0717341	<i>National High-throughput Sequencing Centre,</i> Dinamarca	IlluminaHiSeq	890,670	~98 pb

#### a. Sequenciamento de Nova Geração - NGS

Duas plataformas de sequenciamento NGS foram utilizadas. Para a sequenciamento com sistema *Roche Genome Sequencer FLX* foi utilizado *Titanium Kit* (454 Life Sciences). O sequenciamento foi realizado por Fay et al. (não publicado) em Kew, no Reino Unido. Os autores disponibilizaram as sequências para este estudo. A preparação da biblioteca genômica e o sequenciamento com sistema Illumina HiSeq2000 foram realizados pelo centro nacional dinamarquês de sequenciamento em alta escala (*Danish National High-throughput Sequencing Center*).

### 4.2. ANÁLISE DE DADOS

#### 4.1.1. Sequenciamento Sanger

Eletroferogramas de produtos de PCR de longo alcance (*Long Range PCR*) sequenciados foram editados e montados usando SEQUENCHER versão 4.9 (NEUBIG et al., 2012).

#### 4.1.2. Dados de Sequenciamento de Nova Geração

Os pacotes de *software* FASTQC versão 3 e PRINSEQ versão *little* 0.20.4 foram utilizados para analisar a qualidade de dados de sequência Illumina HiSeq e Roche 454. As *reads* foram filtradas e aparadas (*trimmed*) usando PRINSEQ versão *little* 0.20.4 e Trimmomatic versão 3.2.

#### a. Plastome assembly: Montagem do cpDNA

Para dados de sequenciamento Illumina, sequências-consenso foram geradas através da produção de alinhamentos entre ‘R’ e um genoma de referência, utilizando-se Bowtie2 versão 2.2.3 e Burrows-Wheeler Aligner (BWA) versão 0.7.9a. Para produção de *Contigs* via *de novo* utilizou-se o *software* IOGA versão 1.2 (BAKKER et al., 2015), o qual utiliza SOAPdenovo2 v. R240, Trimmomatic versão 0.32, Assembly Likelihood Evaluation – ALE (*software* de Avaliação da probabilidade de acerto de montagem), samtools versão 0.1.18 e SPAdes *assembler* v. 3.0. Os *softwares* SAMTOOLS versão 0.1.18 e *Integrative Genomic Viewer* (IGV) v. 2.3.32 foram usados para realizar várias etapas de montagem

intermediárias. O *software* GNUplot v. 4.6 foi usado para gerar gráficos de dados de cobertura de montagens.

A Figura 27 mostra etapas gerais do processamento de dados NGS realizados neste trabalho, divididas em três etapas principais. Os passos intermediários utilizados nas etapas um, dois e três são descritos em detalhe nas tabelas 4, 5 e 6 respectivamente. A etapa 1 resume-se na filtragem de *reads* de baixa qualidade e na produção de uma sequência-consenso. Na etapa 2, o *software* IOGA produz *de novo scaffolds* e *contigs*, tendo como produto final uma lista de *assemblies* selecionadas qualitativamente. A etapa 3 integra sequências-consenso e *de novo scaffolds e contigs*.

Montagens do genoma do cloroplasto via *de novo assembly e reference assembly* foram realizadas para *Cypripedium calceolus* e 3 espécimes de *Paphiopedilum barbatum*. Para estudos de genética de populações utilizou-se dados de sequência de cada espécime de *Paphiopedilum barbatum* separadamente. No entanto, devido ao baixo número de *reads* desses espécimes, foi gerado um *dataset* contendo dados de sequência dos três espécimes, que foi utilizado apenas na análise comparativa para marcadores filogenéticos em Cyripedioideae. A percentagem de cpDNA recuperado foi calculado dividindo o tamanho da sequência-consenso (pb) final pelo tamanho (pb) da sequência do cpDNA de *P. longifolium*.

Para dados de sequenciamento Roche 454, os arquivos de consenso foram gerados através de *reference assembly* usando Bowtie2 versão 2.2.3 e Burrows-Wheeler Aligner (BWA) versão 0.7.9a. *De novo contigs* foram gerados usando SOAPdenovo2 versão R240.

#### b. Anotação

Sequências-consenso finais das cinco amostras tiveram seus genomas anotados utilizando Geneious versão 7.1.5, ferramenta *online* cpGAVAS (<http://www.herbalgenomics.org/>, acessado dia 5 de novembro de 2015). Anotação gênica de *Phragmipedium longifolium* foi baseada em *Dendrobium officinale* (número de acesso GenBank KC771275) como genoma de referência. Anotação das demais orquídeas sapatinhos-de-vênus foi baseada na sequência anotada de *Phragmipedium longifolium*. Mapa circular do genoma plastidial de *Phragmipedium longifolium* foi obtido através da ferramenta *online* OGDRAW (<http://ogdraw.mpimp-golm.mpg.de/cgi-bin/ogdraw.pl> acesso dia 5 de novembro de 2015).

c. *Comparação de sequências plastidiais obtidas*

Os alinhamentos foram gerados com MAFFT versão 7.158 e Geneious versão 7.1.5. Usou-se opções padrão para MAFFT, e para alinhamentos gerados com Geneious as seguintes opções foram utilizadas: *Global alignment with free end gaps; Cost matrix of 65% similarity 5/-4; Gap open penalty 12; Gap extension penalty 3; Refinement iterations 2.*

d. *Análise filogenética*

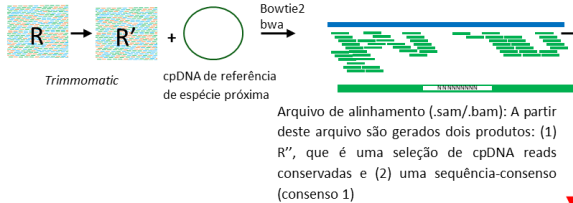
Árvores filogenéticas foram gerados com PAUP 4b10 usando opções de *software: Parsimony, heuristic search starting stepwise, e rooting with an outgroup. Maximum parsimony performing bootstrapping* foi feito com 1.000 repetições, *heuristic search and random seed type.*

e. *Deteção de elementos repetitivos*

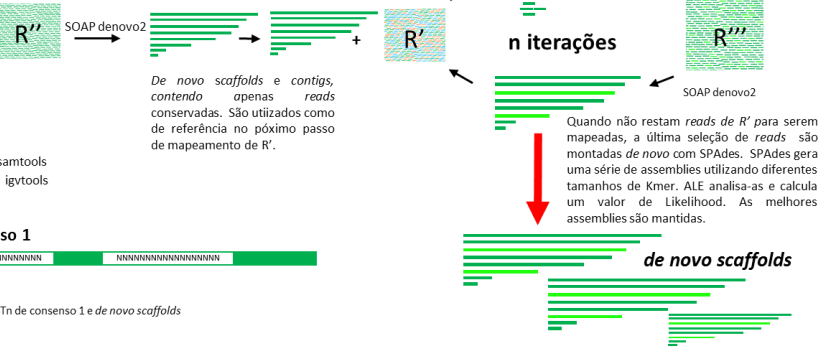
Os elementos repetitivos foram detectados utilizando Phobos versão 3.3.12 como um plugin do Geneious versão 7.1.5, usando o modo "*perfect search*". Esta opção detectou elementos repetitivos de vários comprimentos e de diferentes classes (exemplo, mononucleotídeos, tetranucleotídeos) nas sequências de cloroplastos de três indivíduos de *Paphiopedilum barbatum* alinhados com a sequência anotada de *Phragmipedium longifolium*.

**Figura 27 -** Etapas do processamento de dados NGS realizados neste trabalho.

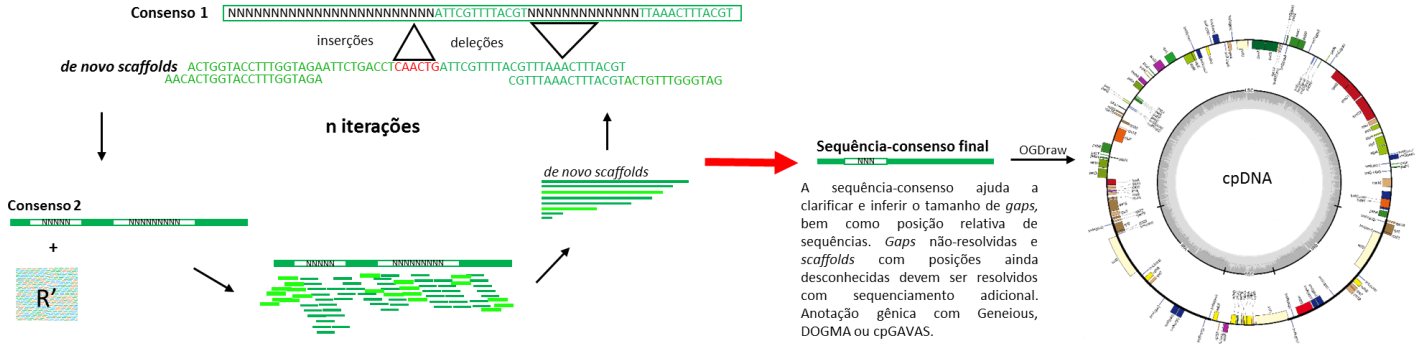
**1. Filtragem de reads de baixa qualidade e produção de uma sequência consenso baseada numa sequência de referência.**



**2. Produção scaffolds via de novo com IOGA**



**3. Combinando seqüências-consenso com de novo scaffolds**





**Tabela 4** - Processamento de dados NGS. Descrição de passos realizados na etapa 1. São ilustrados na Figura 28. **Legenda:** **R**. Readpool contendo todas reads de mtDNA, nDNA e cpDNA geradas por sequenciadores NGS. **R'**. Readpool após etapas de filtragem, somente reads de qualidade dos genomas mtDNA, nDNA e cpDNA. **R''**. Primeira seleção de reads, onde somente seqüências muito conservadas são mantidas.

Passo	Input	Descrição	output
<b>Etapa 1 - Filtragem de reads e produção de Consenso 1</b>			
a. Filtragem de reads	R	<i>Trimmomatic</i> retira reads com baixa qualidade	R'
b. Reference assembly	R' e cpDNA de referência (espécie relacionada)	Bowtie2 ou bwa realizam <i>reference assembly</i> . Reads selecionadas são reads conservadas do cpDNA	arquivo de alinhamento (.sam/.bam)
c. Produção de Sequência-consenso	arquivo de alinhamento (.sam/.bam) produzido no passo anterior	SAMtools ou IGVtools 'colapsam' a sobreposição de reads em uma seqüência linear (.fasta)	Consenso 1 (.fasta)
d. Seleção de reads plastidiais conservadas	arquivo de alinhamento (.sam/.bam) produzido no passo b	Todas as reads que mapearam na seqüência de referência (espécie relacionada) são selecionadas. Este subgrupo de reads é armazenado separadamente, e está representado como R''.	R''

**Tabela 5** - Processamento de dados NGS. Descrição de passos realizados na etapa 2. São ilustrados na Figura 28. **Legenda:** **R.** *Readpool* contendo todas reads de mtDNA, nDNA e cpDNA geradas por sequenciadores NGS. **R’.** *Readpool* após etapas de filtragem, somente reads de qualidade dos genomas mtDNA, nDNA e cpDNA. **R’’.** Primeira seleção de *reads*, onde somente sequências muito conservadas são mantidas.

Passo	Input	Descrição	output
<b>Etapa 2 – Produção de <i>de novo scaffolds</i> com IOGA (Bakker et al. 2005) através de iterações</b>			
a. Filtragem de <i>reads</i>	R	<i>Trimmomatic</i> retira <i>reads</i> com baixa qualidade	R’
b. <i>Reference assembly</i>	R’ e cpDNAs de referência utilizados em IOGA	<i>Bowtie2</i> realiza <i>reference assembly</i> . <i>Reads</i> selecionadas são <i>reads</i> conservadas do cpDNA	arquivo de alinhamento (.sam/.bam)
c. Seleção de <i>reads</i> plastidiais conservadas	arquivo de alinhamento (.sam/.bam) produzido no passo anterior	Todas as <i>reads</i> que mapearam nas sequências de referência são selecionadas. Este subgrupo de <i>reads</i> é armazenado separadamente, e está representado como R’’.	R’’
d. <i>de novo assembly</i>	R’’	<i>SOAPdenovo2</i> utiliza a seleção de <i>reads</i> conservadas para produzir <i>de novo scaffolds</i> e <i>contigs</i>	<i>de novo scaffolds</i> e <i>contigs</i>
e. <i>Reference assembly</i>	<i>de novo scaffolds</i> e <i>contigs</i> produzidos no passo anterior + R’	<i>Bowtie2</i> faz o mapeamento de <i>reads</i> de R’ utilizando <i>de novo scaffolds</i> e <i>contigs</i> . <i>Reads</i> que carregam mais variação são mapeadas.	arquivo de alinhamento (.sam/.bam)
f. Seleção de novas <i>reads</i>	arquivo de alinhamento (.sam/.bam) produzido no passo anterior	Todas as <i>reads</i> que mapearam na sequência de referência ( <i>de novo scaffolds</i> e <i>contigs</i> ) são selecionadas e armazenadas separadamente, formando R’’’. Esta seleção de <i>reads</i> inclui <i>reads</i> altamente conservadas e também <i>reads</i> espécie-alvo específicas.	R’’’

g. <i>de novo assembly</i>	R'''	SOAPdenovo2 utiliza nova seleção de <i>reads</i> (conservadas e não-conservadas) para produzir <i>de novo scaffolds</i> e <i>contigs</i>	<i>de novo scaffolds</i> e <i>contigs</i> alongados
h. Iterações	-	IOGA alterna <i>reference assembly</i> e <i>de novo assembly</i> com intuito de selecionar o máximo de <i>reads</i> de R'.	R''' com máximo de <i>reads</i>
i. <i>de novo assembly</i>	Última seleção de novas <i>reads</i> , isto é, R''' alongada.	SPAdes realiza <i>de novo assembly</i> com a última seleção de <i>reads</i> , utilizando algoritmos e diferentes valores de Kmer (subsequências de <i>reads</i> ). Produz uma lista de <i>assemblies</i> diferentes para cada valor de Kmer utilizado. Cada <i>assembly</i> é um conjunto de <i>de novo scaffolds</i> e <i>contigs</i> .	Lista de <i>assemblies</i> com diferentes Kmer
j. Validação de qualidade das <i>assemblies</i>	Lista de <i>assemblies</i> produzidas por SPAdes	ALE calcula valores de <i>Likelihood</i> para cada <i>assembly</i> e seleciona os valores de Kmer que produziram as melhores	Seleciona as melhores <i>assemblies</i>

**Tabela 6-** Processamento de dados NGS. Descrição de passos realizados na etapa 3. São ilustrados na Figura 28. **Legenda:** **R.** *Readpool* contendo todas reads de mtDNA, nDNA e cpDNA geradas por sequenciadores NGS. **R’.** *Readpool* após etapas de filtragem, somente reads de qualidade dos genomas mtDNA, nDNA e cpDNA. **R’’.** Primeira seleção de *reads*, onde somente sequências muito conservadas são mantidas.

Passo	Input	Descrição	output
<b>Etapa 3 – Integrando sequências-consenso com <i>de novo scaffolds</i> e <i>contigs</i> das melhores <i>assemblies</i></b>			
a. BLASTn de <i>contigs</i> e <i>scaffolds</i> para inferência de sua posição relativa e resolução de <i>gaps</i>	Consenso 1 e <i>de novo scaffolds</i> e <i>contigs</i> de melhor <i>assembly</i>	Realiza-se BLASTn de <i>scaffolds</i> e <i>contigs</i> tanto na plataforma online NCBI quanto pode-se fazer um BLASTn entre Consenso 1 e melhor <i>assembly</i> . O posicionamento relativo de <i>scaffolds</i> pode ser inferido (dependendo do grau de distanciamento do genoma de referência inicial) e também regiões polimórficas podem ser detectadas. Algumas <i>gaps</i> podem ser resolvidas, inserções e deleções detectadas.	arquivo de alinhamento (.sam/.bam)
b. Produção de Consenso 2	arquivo de alinhamento (.sam/.bam) entre Consenso 1 e <i>scaffolds</i> e <i>contigs</i> de melhor <i>assembly</i>	SAMtools ou IGVtools ‘colapsam’ a sobreposição de <i>scaffolds</i> e <i>contigs</i> em uma sequência linear (.fasta), preenchendo lacunas ( <i>gaps</i> ) não-resolvidas com Ns. Esta sequência foi aqui nomeada de Consenso 2.	Consenso 2
c. <i>Reference assembly</i>	Consenso 2 e R’	Nesta etapa Bowtie2 realiza <i>reference assembly</i> de R’ ( <i>Readpool</i> filtrada) com Consenso 2, com intuito de gerar dados de cobertura. Se nesta etapa novas <i>reads</i> forem selecionadas, <i>de novo assembly</i> deve ser realizada.	arquivo de alinhamento (.sam/.bam): Observar dados de cobertura.

d. Seleção de novas <i>reads</i>	arquivo de alinhamento (.sam/.bam) produzido no passo anterior.	Seleção de eventuais novas <i>reads</i> (é possível que novas <i>reads</i> não sejam detectadas no passo anterior).	Novo R'''
e. <i>De novo assembly</i>	Novo R''' produzido no passo anterior.	<i>De novo scaffolds</i> produzidos.	<i>De novo scaffolds</i>
f. iterações	-	Nesta etapa pode-se produzir novas sequências-consenso a partir de <i>scaffolds</i> e <i>contigs</i> alongados, num ciclo iterativo de c, d, e. Quando não há novas informações a serem integradas na sequência-consenso, produz-se a sequência-consenso final. Esta ajuda a inferir tamanhos de gaps e posição relativa de sequências produzidas.	Consenso Final
g. Anotação gênica e produção de mapa cpDNA circular	Sequência-consenso final	Anotação gênica com Geneious, DOGMA ou cpGAVAS pode ser feita. Após anotação gênica, a sequência final anotada pode ser processada por OGDRAW, que a representa de maneira circular, mapeando genes e mostrando conteúdos de CG.	cpDNA circular anotado

## 5. RESULTADOS E DISCUSSÃO

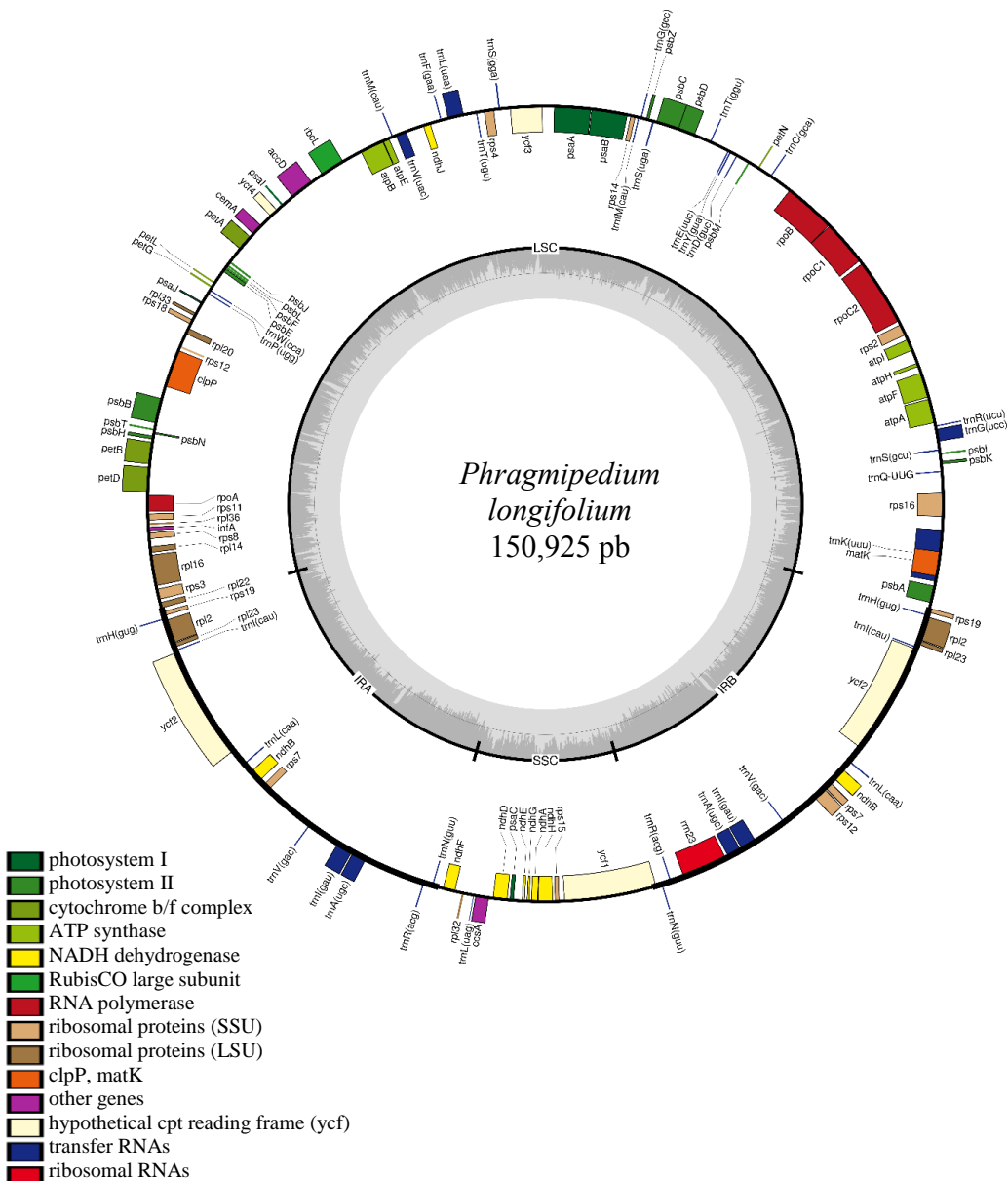
### 5.1. SEQUÊNCIA COMPLETA DO PLASTOMA DE *PHRAGMIPEDIUM LONGIFOLIUM*

O genoma completo do cloroplasto de *Phragmipedium longifolium* constituiu uma molécula circular de 150.925 pb. O conteúdo geral de A + T na sequência foi de 63,9%, comparável aos valores encontrados em outros plastomas de orquídeas, por exemplo, *Phalaenopsis equestris* (148.959 pb; 63,35%), *Phalaenopsis afrodite* (148.964 pb; 62,35%), *Dendrobium officinale* (152.221 pb; 62,5%) e *Oncidium Gower Ramsey* (146.484 pb; 62,7%). Utilizando a ferramenta online BLAST, na plataforma NCBI, a sequência de *Phragmipedium longifolium* mostrou-se mais semelhante ao cpDNA de *Dendrobium officinale* (número de acesso GenBank KC771275), apesar do conteúdo de A + T ser um pouco maior do que o deste último. Um mapa circular do genoma do cloroplasto de *P. longifolium* anotado neste trabalho é mostrado na Figura 28.

Como típico em Angiospermas, a sequência plastidial tem quatro partições estruturais (BENDICH, 2004). O genoma de *P. longifolium* possui duas sequências quase idênticas, as IRs (24.862 pb cada) dividindo as regiões LSC (88.135 pb) e SSC (13.066 pb). A sequência completa possui 107 genes, incluindo 75 genes que codificam para proteínas, 31 genes de RNA transportador e um gene ribossomal. Do total dos genes, 10% contém regiões de introns (17 ao todo, 6 genes de tRNA e 11 codificantes de proteínas). Como encontrado em sequências plastidiais de *Phalaenopsis* sp. (JHENG et al., 2012), os genes *clpP*, *rps12* e *ycf3* contém dois introns cada.

Em termos de conteúdo G + C distribuídos pelo genoma de *P. longifolium*, as regiões codificantes de proteína (no total 68.382 pb; 45,3% do plastoma completo) contém 38% e o conteúdo G + C nas regiões de genes tRNA (2.859 pb; 1,89%) é de 52,2%. O único gene de rRNA presente na sequência (2.807 pb; 1,86%) possui 55,2% de G + C. Sequências não-codificantes (76.877 pb; 50,93%), incluindo pseudogenes, espaçadores e introns, apresentaram 34,2% de conteúdo G + C. As regiões plastidiais IRs mostraram um maior conteúdo geral de G + C quando comparadas às regiões de cópias simples (Figura 28). Elas apresentaram um teor de G + C de 43,0%, enquanto que as regiões SSC e LSC mostram 29,2% e 33,5%, respectivamente.

**Figura 28** - Mapa circular do plastoma de *Phragmipedium longifolium*, anotado neste estudo. Genes interiores e exteriores ao círculo são transcritos no sentido horário e anti-horário respectivamente. Grupos gênicos são indicados de acordo com as cores. Limites das regiões LSC, IRa, IRb e SSC são mostrados no círculo interior. O gráfico interno mostra conteúdo de G + C ao longo da sequência.



## 5.2. SEQUÊNCIAS PLASTIDIAIS OBTIDAS

Sequências plastidiais foram montadas a partir de uma amostra sequenciada com Roche 454 e de três amostras sequenciadas com Illumina HiSeq2000. Um *dataset* combinando dados de sequenciamento Illumina de três espécimes de *Paphiopedilum barbatum* também foi processado e uma *assembly* (ou ‘montagem’) também foi gerada. Os detalhes das montagens e o tamanho (pb) dos arquivos de consenso-final e da melhor *assembly* produzida por IOGA são mostrados na Tabela 7. A profundidade de leitura (dados de cobertura) de cada conjunto é mostrada na Figura 29.

Para o *dataset* contendo sequências de todos os três indivíduos de *Paphiopedilum barbatum*, as etapas 1, 2 e 3, descritas nas Tabelas 4, 5 e 6, foram realizadas. A montagem final dos dados combinados de *Paphiopedilum barbatum* contiveram 143.480 pb, onde 0,73% (1.056 posições) consistiram de dados em falta ou ambiguidades. Três grandes *gaps* (entre 200 e 450 pb) e 10 pequenas *gaps* (de 1 a 10 pb) não puderam ser resolvidas sem sequenciamento adicional. Levando em consideração que o tamanho do cpDNA de *Phragmipedium longifolium* é 150.925 pb, pode ainda estar faltando nesta montagem cerca de 7.000 pb. Esta questão só poderia ser resolvida com sequenciamento adicional, que não foi realizada devido à restrição de tempo. No entanto, um total de 1.426.439 *reads* (5,07% do total de 28.092.223) foram mapeadas na montagem final, recuperando cerca de 95,07% do cpDNA. A profundidade média de *reads* foi de 11,6x. O conteúdo total de A + T deste conjunto foi de 63,9%, semelhante ao encontrado em *P. longifolium* (63,9%) e comparável ao conteúdo de A + T plastidial de outras orquídeas.

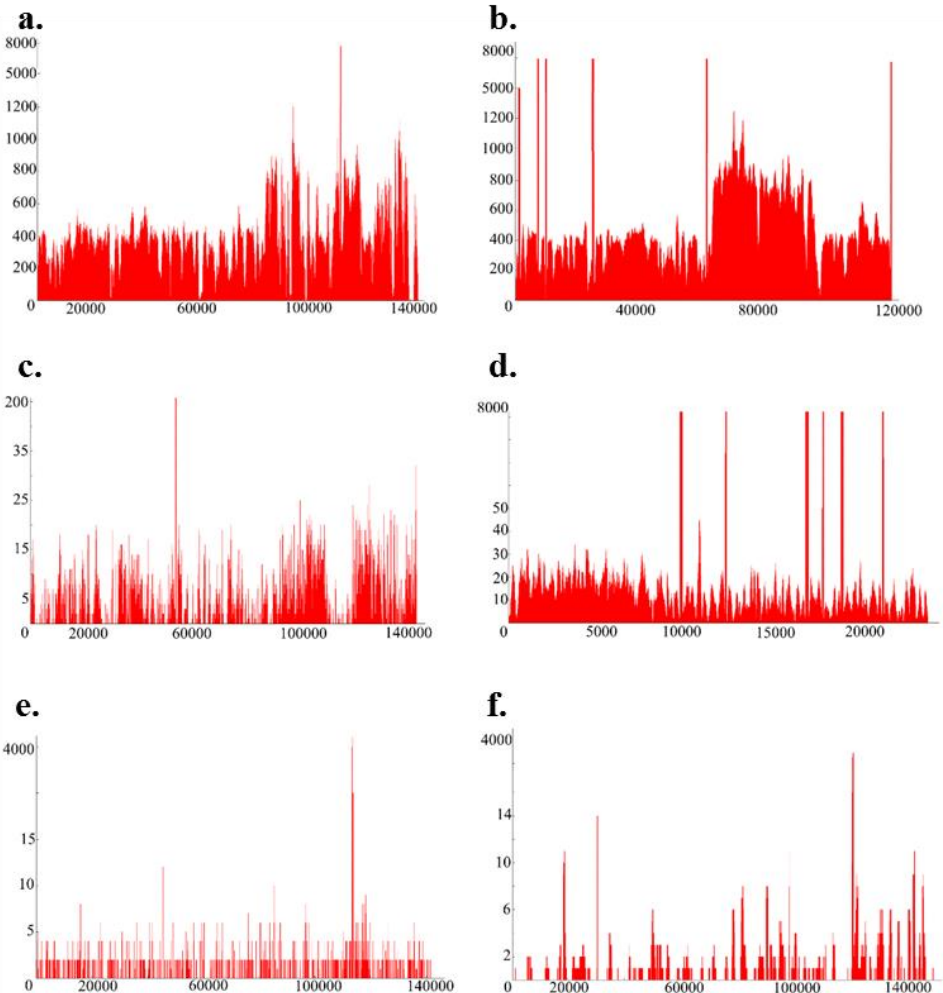
A montagem final (*final assembly*) dos dados combinados de *Paphiopedilum barbatum* foi utilizada como sequência de referência para realizar *reference assembly* para espécimes frescos e de herbário de *P. barbatum* separadamente. Sequências-consenso foram geradas sem utilizar *de novo scaffolds* e *contigs*. Para a amostra fresca de *Paphiopedilum barbatum* 93,45% do cpDNA foi recuperado, e 868.652 *reads* (5,75% dos 15.097.287) foram mapeadas na sequência de referência. Um arquivo de consenso foi gerado, de comprimento 142,936 pb, onde 1,33% (1.899 pb) consistiu de dados em falta e ambiguidades. O conteúdo A + T desta sequência foi de 36,2%, e a profundidade média de leitura foi 6,16x. Dados de sequenciamento desta amostra fresca também foram processados no software IOGA, tendo sido produzido *de novo scaffolds* e também o relatório das melhores *assemblies*. No entanto, para



**Tabela 7:** Sequências de cpDNA montadas a partir de dados NGS. *Reference assembly* gerou Consenso 1 e IOGA (BAKKER et al., 2015) gerou *de novo scaffolds*. A montagem final do cpDNA de *Phragmipedium longifolium* e a montagem final dos dados combinados de 3 indivíduos de *Paphiopedilum barbatum* foram utilizados como referência.

<b>Amostra</b>	<b>Plat.</b>	<b># total de reads</b>	<b>Tamanho de reads (pb)</b>	<b>Consenso 1 (pb)</b>	<b><i>de novo scaffolds</i> (pb)</b>	<b>Consenso Final (pb)</b>	<b># reads mapeadas</b>	<b>% cpDNA recuperado</b>
<i>Cypripedium calceolus</i>	Roche 454	34.048	300 a 1500	61.765	-	61.765	325	41,02
<i>P. barbatum</i> , dados combinados	Illumina HiSeq	28.092.223	~99	143.480	125.336	143.480	1.426.439	95,07
<i>Paphiopedilum barbatum</i> (2013)	Illumina HiSeq	15.097.287	~100	141.037	125.813	141.037	868.652	93,45
<i>Paphiopedilum barbatum</i> (1970)	Illumina HiSeq	890.670	~98	~32.456	164	32.456	6.802	21,50
<i>Paphiopedilum barbatum</i> (1968)	Illumina HiSeq	12.104.266	~98	89.825	24.178	89.825	152.87	59,52

**Figura 29** - **a-e**. Dados de cobertura de montagens do cpDNA de espécimes de *P. barbatum*. Os gráficos mostram a posição das bases nucleotídicas (pb) no eixo 'x' e o número de sobreposições de reads (reads que mapearam numa determinada posição) no eixo y. **a-b**. Amostra de *P. barbatum* coletada em 2013. **a**. *Reference assembly* e valores de cobertura. **b**. Valores de cobertura para *de novo scaffolds* (observar que um *scaffold* contém a sequência das regiões invertidas e portanto tem o dobro de reads mapeadas). **c-d**. Espécime de herbário coletado em 1968. **c**. *Reference assembly*. **d**. Valores de cobertura de *de novo scaffolds*. **e**. Valores de cobertura de *reference assembly* de *P. barbatum* coletado em 1970. **f**. Valores de cobertura de *reference assembly* de *Cyripedium calceolus*.



a finalidade deste estudo, foi suficiente trabalhar com os arquivos de consenso.

Para o espécime de herbário com 44 anos de idade, coletados em 1970, 21,51% do cpDNA foi recuperado. Para esta amostra, duas montagens de referência foram realizadas usando tanto *Paphiopedilum barbatum* como *Phragmipedium longifolium* como referência, os quais foram fundidos em um arquivo de consenso. No total 6.802 *reads* abrangendo 32.456 bases foram mapeadas em um arquivo de consenso de 141.711 pb de comprimento onde 77,1% foram dados em falta. O conteúdo total de A + T foi de 65,3% quando desconsideradas as bases ambíguas. O espécime de herbário com 46 anos de idade, coletado em 1968, teve 59,52% de seu cpDNA recuperado. No total 152.870 *reads* (1,26% dos 12.104.266pb) foram mapeadas no genoma de referência, totalizando 89.825 nucleotídeos mapeados. O arquivo de consenso possui 142.936 pb de comprimento, onde 37,3% (53.316 pb) foram dados em falta ou ambiguidades. A média de A + T foi de 59,0% sem levar em conta as bases ambíguas.

Para *Cypripedium calceolus*, 41,02% das sequências de cpDNA foram recuperadas. No total, 325 *reads* de comprimentos variáveis foram mapeadas para o cpDNA de *Phragmipedium longifolium*. O arquivo de consenso consiste em uma sequência de 150.569 pb onde 58,0% (88.804 pb) foram dados em falta. O teor médio de A + T foi de 63,0%.

### 5.3. MARCADORES FILOGENÉTICOS PUTATIVOS PARA ORQUÍDEAS SAPATINHOS-DE-VÊNUS

As sequências do genoma plastidial de três orquídeas sapatinhos-de-vênus foram comparadas e o nível de variação entre regiões gênicas e intergênicas localizadas nas IRs, SSC e LSC foi inferido. O alinhamento consistiu na sequência de cpDNA de *Phragmipedium longifolium* anotada, e nas sequências-consenso finais de *Paphiopedilum barbatum*, de *Cypripedium calceolus* e *Dendrobium officinale* (número de acesso GenBank KC771275). Não foi possível realizar análise filogenética posto que tivemos apenas três representantes de Cypripedioideae neste alinhamento. Entretanto, o nível de variação das regiões do cpDNA entre os 4 espécimes foi medido (Tabela 8).

Entre os 29 marcadores putativos detectados para orquídeas sapatinhos-de-vênus, o nível de variação oscilou de 3,09% a 57,54% (bases nucleotídicas variáveis / número total de bases comparadas). Isto confirma que, apesar de ter uma estrutura geral muito conservada, o genoma do cloroplasto de orquídeas carrega regiões informativas

**Tabela 8:** Marcadores filogenéticos putativos para orquídeas sapatinhos-de-vênus detectados neste estudo. Marcadores clássicos são indicados em negrito.

Região	Start	End	Tamanho (pb)	# posições idênticas	# posições polimórficas	% variação
<i>trnL-trnF</i>	58064	59755	1691	718	973	57.54
<i>atpB-rbcL</i>	65930	66925	995	451	544	54.67
<i>matk</i>	.	.	1,347	820	527	39.12
<i>accD</i>	69937	70839	902	643	259	28.71
<i>ndhJ</i>	60449	60976	527	397	130	24.67
<i>ycf1</i>	149236	152981	3745	2833	912	24.35
<i>rbcL</i>	67445	69023	1578	1206	372	23.57
<i>clpP</i>	84478	85437	959	775	184	19.19
<i>rpoC1</i>	24445	27309	2864	2,437	427	14.91
<i>ycf3</i>	53549	54531	982	846	136	13.85
<i>rps16</i>	6736	7629	893	794	99	11.09
<i>petL-trnP</i>	79490	80349	859	770	89	10.36
<i>ycf1</i>	153112	154723	1611	1455	156	9.68
<i>rpl16-rps3</i>	100374	101141	767	693	74	9.65
<i>psbB</i>	88838	89466	628	577	51	8.12
<i>ycf2</i>	106239	107212	973	903	70	7.19
<i>rpl22-rpl2</i>	101779	103286	1507	1399	108	7.17
<i>atpA-atpF</i>	13320	14882	1562	1,451	111	7.11
<i>petB-petD</i>	91491	92635	1144	1068	76	6.64
<i>rpoA-rps8</i>	94873	97088	2215	2073	142	6.41
<i>rps12</i>	116645	117449	804	754	50	6.22
<i>ycf2</i>	107415	108476	1061	997	64	6.03
<i>psbD</i>	42650	43544	894	844	50	5.59
<i>ndhB-rps7</i>	115499	116472	973	920	53	5.45
<i>rpl2</i>	103363	104133	770	730	40	5.19
<i>atpE</i>	64016	64428	412	391	21	5.09
<i>ycf2</i>	104879	105700	821	787	34	4.14
<i>ccsA</i>	141069	141901	832	802	30	3.61
<i>trnV</i>	62546	63226	680	659	21	3.09

filogeneticamente, como visto em estudos anteriores (CHANG et al., 2006; WU et al., 2010; JHENG et al., 2012; PAN et al., 2012; YANG et al., 2013; LIN et al., 2015).

No total 19 genes conservados codificantes de proteínas foram detectados, 10 deles localizados na região do LSC, 5 nos IRS e 3 na região SSC. Nove seqüências intergênicas (por exemplo, *trnL-trnF*) mostraram alto nível de variação (foram consideradas seqüências de genes e as seqüências que os conectam), 7 localizadas no LSC e 2 no IRS. Um pseudogene foi detectado como uma região promissora (*ndhJ*), localizado no LSC. Os resultados não demonstram níveis mais elevados de variação em espaçadores intergênicos, não estando de acordo com a suposição comum que regiões codantes do DNA geralmente evoluem mais lentamente que regiões não-codantes (DOORDUIN et al., 2011). Isto poderia ser um artefato devido ao número restrito de seqüências comparadas.

O presente trabalho encontrou diferenças significativas de níveis de variação de diferentes regiões plastidiais (LSC, SSC e IRS). A região LSC apresentou maior nível de variação em relação a outras regiões plastidiais, apresentando 65,52% das regiões promissoras, seguido por 24,15% e 10,35% para as regiões IRS e SSC, respectivamente. Yang et al. (2013) também encontraram diferenças significativas nos níveis de variação entre regiões plastidiais de orquídeas. No entanto a região SSC demonstrou ser a mais variante, com 3,5% de caracteres filogeneticamente informativos, sendo as regiões IRS as mais conservadas, com 0,9%.

Altos níveis de variação foram detectados em seqüências intergênicas (gene-intergene-gene). A região *trnL-trnF* (1,691pb) mostrou o maior nível de variabilidade, com 57,54% de bases polimórficas, seguido pela seqüência intergênica *atpB-rbcL* (995pb), com 54,67% de variação. Estas regiões estão localizadas na região plastidial LSC. A escolha de *trnL-trnF* como marcador molecular em Orchidaceae é também suportada por estudos anteriores que demonstraram sua eficácia. Van den Berg et al. (2005) encontraram 49,1% de variação na região *trnL-trnF* ao comparar orquídeas da subtribo Laeliinae (Epidendroideae). Realizando análise filogenética, os autores encontraram 27,6% de caracteres filogeneticamente informativos de um total de 1622pb. Van den Berg et al. (2005) ressaltam que para uma região não-codificante este é um nível comum de variação, como também visto em Bakker et al. (1999). Seqüências intergênicas têm sido bem sucedidas para esclarecer as relações filogenéticas a nível de gênero em Orchidaceae

(como YANG et al., 2013, gênero *Cymbidium*) e a nível de subtribo (como PAN et al., 2012, Oncidiinae).

Por outro lado, marcadores clássicos também mostraram alta variação entre as amostras (em negrito na Tabela 8). A terceira sequência mais variável foi o marcador clássico *matK* (1,347pb), com 39,12% de bases variáveis, seguido por *accD* (902pb), com 28,71%. Ambos são sequências codificantes de proteínas localizadas na região LSC. No entanto, a utilização de marcadores clássicos, em estudos filogenéticos com orquídeas sapatinho-de-vênus, foi insuficiente para gerar boa resolução abaixo de nível de gênero (COX et al., 1997; CHOCHAI et al., 2012; GUO et al., 2012) e a concatenação com regiões não-codificantes informativas de rápida evolução se faz necessária. Wu et al. (2010) obtiveram sucesso em distinguir 15 variedades de *Oncidium* sp. utilizando a sequência codificante de *matK*, juntamente com sete sequências não-codificantes plastidiais. Os autores relatam que a sequência codificante de proteína *matK* mostrou ser informativa a nível de espécies. Van den Berg et al. (2009), estudando orquídeas da subtribo Laeliinae, também concluíram que o gene *matK* é frequentemente útil em todos os níveis taxonômicos.

Variação em pseudogenes *ndh* foi detectada durante a análise do alinhamento produzido no presente trabalho, entretanto, não havia sequências recuperadas suficientes para comparar todas as subunidades *ndh* entre as três orquídeas sapatinhos-de-vênus. As subunidades *ndhJ* e *ndhB* puderam ser detectadas e foram analisadas entre as amostras, mostrando 24,67% e 87,01% de variação respectivamente. A subunidade *ndhJ* foi a quinta região mais variável. O alto valor de *ndhB* é devido a uma inserção de 1.469pb em *Paphiopedilum barbatum*, enquanto em *Cypripedium calceolus* e *Phragmipedium longifolium* o gene se encontrava truncado (ou seja, havia *codons* interrompidos na sequência), por isto esta sequência não foi incluída na Tabela 8. A subunidade *ndhK* não estava presente em nenhuma das amostras. Os pseudogenes *ndh* estavam dispersos em todas as regiões de cpDNA, sendo que, em *Phragmipedium longifolium*, o *ndhJ* do cpDNA está localizado na região LSC, enquanto o *ndhB* está localizado no IRs e o *ndhF*, D, E, G, A, H estão localizados no SSC (Figura 28). Estes resultados estão de acordo com o que tem sido relatado para Orchidaceae, onde perdas e truncagens das subunidades *ndh* são bastante comuns (CHANG et al., 2006; GUO et al., 2012; JHENG et al., 2012; PAN et al., 2012; YANG et al., 2013; LIN et al., 2015). A razão para deleções e perda de função de subunidades *ndh* pode ter sido devido ao fato de que as cópias *ndh* de plastídios ancestrais tenham sido transferidas para o núcleo.

Há onze subunidades compondo a família gênica *ndh*, que codificam para o complexo protéico NADH-desidrogenase. Este complexo protéico parece participar em processos de adaptação da resposta fotosintética a estresse ambiental (RUHLMAN et al., 2015). Lin et al. (2015) estudaram padrões evolutivos de subunidades *ndh* em genoma plastidial de orquídeas. Os autores não encontraram, no entanto, diferenças significativas entre orquídeas ‘*ndh*-deletadas’ and ‘*ndh*-completas’, em termos de biogeografia e condições de cultivo. Os autores compararam sequências plastidiais inteiras e transcriptomas, com amostragem cobrindo as cinco subfamílias de Orchidaceae, porém nenhuma correlação de eventos de truncagens e deleções de subunidades *ndh* com a filogenia conhecida da família foi verificada. Apesar disto, neste estudo encontraram-se altos níveis de variação na sequência *ndhJ*, sendo que a sequência estava presente em todos os espécimes. O marcador deve ser testado e a porcentagem de caracteres informativos devem ser calculadas. A comparação de árvores filogenéticas geradas com concatenações robustas de marcadores sugeridos na tabela 8 com árvores filogenéticas geradas com *ndhJ* poderia trazer informações sobre a evolução de *ndh*.

Marcadores filogenéticos putativos propostos neste estudo, tais como *trnL-trnF*, *atpB-rbcL*, *matk*, *accD*, *ycf1*, *petL-trnP*, *rpl16-rps3*, *rpl22-rpl2*, *atpA-atpF*, *petB-petD*, *rpoA-rps8*, são bons candidatos a marcadores para filogenética de Cyripedioideae. A porcentagem de caracteres filogeneticamente informativos das sequências, índice de consistência e índice de retenção ainda precisam ser inferidos para confirmar a eficácia destes marcadores. No presente estudo, isto não pode ser alcançado, pois não havia terminais suficientes (isto é, número de espécimes comparados) para representar as orquídeas sapatinhos-de-vênus. Entretanto, com base em resultados de estudos filogenéticos moleculares dentro de Epidendroideae e na comparação de regiões cpDNA entre orquídeas sapatinhos-de-vênus aqui apresentadas, novos marcadores para a subfamília Cyripedioideae poderão ser utilizados em estudos filogenéticos futuros.

#### 5.4. SSRs PUTATIVOS PARA *PAPHIOPEDILUM BARBATUM*

O alinhamento incluiu as três sequências-consenso de indivíduos de *P. barbatum* produzidas neste trabalho e a sequência plastidial de *Phragmipedium longifolium*. Sequências repetitivas contendo variação entre os espécimes foram detectadas, tanto em regiões codificadoras de proteínas como em regiões não-codificantes. Todas as regiões do genoma

plastidial (LSC, SSC, IRS) apresentaram pelo menos uma região repetitiva. Estas regiões são conhecidas como repetições de sequências simples (*Simple Sequence Repeat* - SSR), ou repetições curtas em *tandem* (*Short Tandem Repeats* - STRs). Dentro da classe de SSRs estão os microssatélites, que são regiões de DNA que contêm unidades de repetição em *tandem* de 1-6pb de tamanho (PANDEY e SHARMA, 2012).

Regiões SSRs variáveis entre espécimes de *P. barbatum* são descritas na Tabela 9. A detecção de microssatélites e outros SSRs foi restrita, pois os espécimes de herbário tiveram apenas 59,2 e 21,5% de seu genoma recuperado. No entanto, puderam ser detectados 54 SSRs putativos, sendo a maior parte repetições de hexanucleotídeos (29,6%), seguido por mononucleotídeos (20,4%), trinucleotídeos (14,8%), pentanucleotídeos (13,0%), dinucleotídeos (7,4%), tetranucleotídeos (7,4%), heptanucleotídeos (3,7%) e nonanucleotídeos (3,7%). Comprimento dos alelos variou de 3 a 18 nucleotídeos. SSRs foram localizadas principalmente na região de LSC (77,8%), seguido pela região de SSC (16,7%).

Para este estudo, 21 SSRs foram escolhidas com base na variabilidade entre os três espécimes de *P. barbatum* e na disponibilidade de regiões recuperadas (mostradas em negrito no Tabela 9). Esta seleção de SSRs poderá ser testada e utilizada em estudos populacionais futuros com a espécie. Onze SSRs foram detectados na região LSC (52,4%). Nove SSRs foram detectados nas regiões IRs (42,8%), dos quais 5 deles estavam presentes apenas no IRb, e 4 estavam em ambos IRs. O SSC apresentou apenas um SSR. Sequências de codificação de proteína e sequências não-codificantes, apresentaram 10 (47,6%) e 11 (52,4%) SSRs, respectivamente. Era esperado que sequências não-codificantes possuíssem maior variação, mas os resultados mostraram uma distribuição aparentemente homogênea. Esta distribuição igual de variação poderia ser um artefato, pois várias regiões não puderam ser comparadas.

Dentre os 21 SSRs putativos, repetições de hexanucleotídeos foram as mais abundantes (38,1%), seguido de mononucleotídeos (19,0%), trinucleotídeos (19,0%) e tetranucleotídeos (14,3%). Repetições de nonanucleotídeos (4,8%) e dinucleotídeos (4,8%) foram menos frequentes. Não foram encontrados pentanucleotídeos ou heptanucleotídeos para os três espécimes. Configurações de SSRs de mononucleotídeos apresentaram maior variação, às vezes apresentando três alelos e de tamanho entre 2pb a 16pb. Porém a utilização de SSRs mononucleotídicos não é aconselhada devido à dificuldade de detecção



dos variantes. O número de alelos, ou seja, de repetições em casa SSR, foi expresso em valor aproximado, devido a possíveis erros durante a realização de reações PCR-amplificadoras, bem como a erros em etapas de montagem de DNA, como visto em Doorduyn et al. (2011).

**Tabela 9:** Marcadores para genética de populações de *P. barbatum*. Informações de regiões polimórficas detectadas em indivíduos de *Paphiopedilum barbatum* e um indivíduo de *Phragmipedium longifolium*. ‘CDS’ indica regiões codantes (*Coding Sequence*). Classes de sequências repetitivas SSRs estão abreviadas. Regiões que estavam presentes e eram polimórficas entre espécimes de *P. barbatum* são mostradas em negrito.

Un. de repetição	Classe de SSR	Intervalo	$\Delta$ (pb)	Região	<i>P. longifolium</i> , material fresco, Equador	<i>P. barbatum</i> , material fresco, s.d.	<i>P. barbatum</i> , herbário, Malásia, Terengganu	<i>P. barbatum</i> , herbário, Malásia, Penang Hill
A	mono	15,890 -> 15,897	5-8	<i>atpF</i> CDS*	(A) <sub>8</sub>	(A) <sub>8</sub>	N	(A) <sub>5</sub>
T	mono	23,330 -> 23,336	3-7	<i>rpoC2</i> CDS	(T) <sub>7</sub>	(T) <sub>6</sub>	N	(T) <sub>3</sub>
T	mono	19,709 -> 19,695	10-15	<i>rps2-rpoC2</i>	(T) <sub>10</sub>	(T) <sub>15</sub>	N	(T) <sub>15</sub>
C	mono	49,943 -> 49,949	4-7	<i>psaA</i> CDS	(C) <sub>4</sub>	(C) <sub>4</sub>	N	(C) <sub>7</sub>
C	<b>mono</b>	<b>116,856 -&gt; 116,862</b>	<b>3-7</b>	<b><i>rps12</i> (intron)</b>	<b>(C)<sub>3</sub></b>	<b>(C)<sub>7</sub></b>	<b>(C)<sub>3</sub></b>	<b>(C)<sub>7</sub></b>
A	mono	151,497 -> 151,503	6-7	<i>ycf1</i> CDS	(A) <sub>7</sub>	(A) <sub>7</sub>	N	(A) <sub>6</sub>
T	mono	153,767 -> 153,773	6-7	<i>ycf1</i> CDS	(T) <sub>6</sub>	(T) <sub>6</sub>	N	(T) <sub>7</sub>
T	<b>mono</b>	<b>155,980 -&gt; 155,999</b>	<b>5-16</b>	<b><i>trnR-rrn5</i></b>	<b>(T)<sub>5</sub></b>	<b>(T)<sub>16</sub></b>	<b>(T)<sub>5</sub></b>	<b>(T)<sub>16</sub></b>
A	<b>mono</b>	<b>156,018 -&gt; 156,024</b>	<b>2-7</b>	<b><i>trnR-rrn5</i></b>	<b>(A)<sub>7</sub></b>	<b>(A)<sub>2</sub></b>	<b>(A)<sub>7</sub></b>	<b>(A)<sub>2</sub></b>
A	<b>mono</b>	<b>161,205 -&gt; 161,212</b>	<b>6-8</b>	<b><i>trnI</i> (GAU) (intron)</b>	<b>(A)<sub>7</sub></b>	<b>(A)<sub>6</sub></b>	<b>(A)<sub>8</sub></b>	<b>(A)<sub>6</sub></b>
G	mono	166,994 -> 167,000	3-7	<i>rps12</i> (intron)	(G) <sub>3</sub>	(G) <sub>3</sub>	N	(G) <sub>7</sub>
AT	di	23,192 -> 23,199	4-8	<i>rpoC2</i> CDS	(AT) <sub>4</sub>	(AT) <sub>4</sub>	N	(AT) <sub>2</sub>
AT	di	23,281 -> 23,291	4-10	<i>rpoC2</i> CDS	(AT) <sub>4</sub>	(AT) <sub>4</sub>	N	(AT) <sub>5</sub>
AG	<b>di</b>	<b>68,001 -&gt; 68,008</b>	<b>4-8</b>	<b><i>rbcL</i> CDS</b>	<b>(AG)<sub>2</sub></b>	<b>(AG)<sub>2</sub></b>	<b>(AG)<sub>2</sub></b>	<b>(AG)<sub>4</sub></b>
TC	di	74,513 -> 74,522	8-10	<i>cemA</i> CDS	(TC) <sub>4</sub>	(TC) <sub>4</sub>	N	(TC) <sub>5</sub>
TTA	tri	<b>35,180 -&gt; 35,188</b>	<b>6-9</b>	<b><i>trnC-petN</i></b>	<b>(TTA)<sub>2</sub></b>	<b>(TTA)<sub>2</sub></b>	<b>(TTA)<sub>2</sub></b>	<b>(TTA)<sub>3</sub></b>

TTG	tri	49,965 -> 49,973	3-9	<i>psaA</i> CDS	(TTG) <sub>3</sub>	(TTG) <sub>3</sub>	N	(TTG) <sub>1</sub>
GCA	tri	<b>50,258 -&gt; 50,266</b>	6-9	<i>psaA</i> CDS	(GCA) <sub>2</sub>	(GCA) <sub>2</sub>	(GCA) <sub>3</sub>	(GCA) <sub>3</sub>
AAG	tri	114,174 -> 114,182	0-9	<i>ndhB</i> CDS	(AAG) <sub>0</sub>	(AAG) <sub>2</sub>	(AAG) <sub>2</sub>	(AAG) <sub>3</sub>
CTT	tri	114,395 -> 114,403	0-9	<i>ndhB</i> CDS	(CTT) <sub>0</sub>	(CTT) <sub>2</sub>	(CTT) <sub>2</sub>	(CTT) <sub>3</sub>
TCC	tri	118,235 -> 118,245	3-9	<i>rps12-trnV</i>	(TCC) <sub>3</sub>	N	(TCC) <sub>3</sub>	(TCC) <sub>1</sub>
TTA	tri	142,782 -> 142,790	6-9	<i>ndhD</i> gene	(TTA) <sub>2</sub>	(TTA) <sub>2</sub>	N	(TTA) <sub>3</sub>
GTT	tri	163,346 -> 163,354	6-9	<i>rrn16</i> gene	(GTT) <sub>3</sub>	(GTT) <sub>3</sub>	N	(GTT) <sub>2</sub>
AAAT	tetra	48,755 -> 48,764	4-8	<i>psaB</i> CDS	(AAAT) <sub>1</sub>	(AAAT) <sub>1</sub>	(AAAT) <sub>1</sub>	(AAAT) <sub>2</sub>
TTTC	tetra	97,655 -> 97,665	4-8	<i>rpl14</i> CDS	(TTTC) <sub>1</sub>	(TTTC) <sub>1</sub>	(TTTC) <sub>2</sub>	(TTTC) <sub>1</sub>
AACT	tetra	163,817 -> 163,826	4-8	<i>trnV-rps12</i> ( <i>rps12</i> intron)	(AACT) <sub>2</sub>	(AACT) <sub>1</sub>	(AACT) <sub>2</sub>	(AACT) <sub>1</sub>
AAAC	tetra	167,814 -> 167,824	4-8	<i>rps7-ndhB</i>	(AAAC) <sub>2</sub>	(AAAC) <sub>2</sub>	N	(AAAC) <sub>1</sub>
TTTCA	penta	29,602 -> 29,612	5-10	<i>rpoB</i> CDS	(TTTCA) <sub>1</sub>	(TTTCA) <sub>2</sub>	N	(TTTCA) <sub>1</sub>
TTGGC	penta	32,877 -> 32,887	5-10	<i>rpoB-trnC</i>	(TTGGC) <sub>2</sub>	(TTGGC) <sub>1</sub>	(TTGGC) <sub>1</sub>	N
AAGCC	penta	45,155 -> 45,166	0-10	<i>psbC-</i> <i>trnS(UGA)</i>	(AGCC) <sub>0</sub>	(AGCC) <sub>2</sub>	N	(AGCC) <sub>1</sub>
GGATT	penta	45,669 -> 45,678	5-10	<i>psbZ</i> CDS	(GGATT) <sub>2</sub>	(GGATT) <sub>2</sub>	N	(GGATT) <sub>1</sub>
TAGCT	penta	151,629 -> 151,639	5-10	<i>ycf1</i> CDS	(TAGCT) <sub>1</sub>	(TAGCT) <sub>1</sub>	N	(TAGCT) <sub>2</sub>
TGGTA	penta	179,248 -> 179,258	5-10	<i>trnI-rpl23</i>	(TGGTA) <sub>2</sub>	N	(TGGTA) <sub>2</sub>	(TGGTA) <sub>1</sub>
AAAGT	penta	124,450 -> 124,460	0-12	<i>trnA-rrn23</i>	(AAAGT) <sub>0</sub>	(AAAGT) <sub>2</sub>	(AAAGT) <sub>0</sub>	(AAAGT) <sub>2</sub>
AACTTC	hexa	251 -> 262	6-12	<i>psbA</i> CDS	(GTTGAA) <sub>1</sub>	(GTTGAA) <sub>1</sub>	N	(GTTGAA) <sub>2</sub>
TTGATC	hexa	22,555 -> 22,568	6-12	<i>rpoC2</i> CDS	(TTGATC) <sub>2</sub>	(TTGATC) <sub>2</sub>	N	(TTGATC) <sub>2</sub>
AATTCC	hexa	23,236 -> 23,247	6-12	<i>rpoC2</i> CDS	(AATTCC) <sub>2</sub>	(AATTCC) <sub>2</sub>	N	(AATTCC) <sub>1</sub>
AAAGC C	hexa	27,340 -> 27,351	6-12	<i>rpoB</i> CDS	(AAAGCC) <sub>1</sub>	(AAAGCC) <sub>2</sub>	(AAAGCC) <sub>1</sub>	(AAAGCC) <sub>2</sub>
TGAAG T	hexa	29,414 -> 29,426	6-12	<i>rpoB</i> CDS	(TGAAGT) <sub>1</sub>	(TGAAGT) <sub>2</sub>	N	(TGAAGT) <sub>1</sub>
TAGCT G	hexa	43,686 -> 43,697	0-12	<i>psbC</i> CDS	(TAGCTG) <sub>0</sub>	(TAGCTG) <sub>1</sub>	(TAGCTG) <sub>2</sub>	(TAGCTG) <sub>2</sub>
TGCAG C	hexa	44,971 -> 44,982	6-12	<i>psbC</i> CDS	(TGCAGC) <sub>2</sub>	(TGCAGC) <sub>2</sub>	N	(TGCAGC) <sub>1</sub>
TGGCC C	hexa	50,465 -> 50,476	6-12	<i>psaA</i> CDS	(TGGCCC) <sub>2</sub>	(TGGCCC) <sub>2</sub>	N	(TGGCCC) <sub>1</sub>

TCCCGA	hexa	54,025 -> 54,037	6-12	<i>ycf3</i> gene	(TCCCGA) <sub>1</sub>	(TCCCGA) <sub>1</sub>	N	(TCCCGA) <sub>2</sub>
CCTTCA	hexa	54,259 -> 54,271	6-12	<i>ycf3</i> gene	(CCTTCA) <sub>1</sub>	(CCTTCA) <sub>2</sub>	(CCTTCA) <sub>2</sub>	(CCTTCA) <sub>1</sub>
AGATTG	hexa	63,301 -> 63,313	0-12	<i>psbC-atpB</i>	(AGATTG) <sub>0</sub>	(AGATTG) <sub>1</sub>	(AGATTG) <sub>2</sub>	(AGATTG) <sub>1</sub>
TTCTTA	hexa	115,921 -> 115,932	6-12	<i>ndhB-rps7</i>	(TTCTTA) <sub>2</sub>	(TTCTTA) <sub>2</sub>	(TTCTTA) <sub>1</sub>	(TTCTTA) <sub>2</sub>
TTTCACT	hexa	116,786 -> 116,798	6-12	<i>rps12 (intron)</i>	(TTTCACT) <sub>2</sub>	(TTTCACT) <sub>2</sub>	(TTTCACT) <sub>1</sub>	(TTTCACT) <sub>2</sub>
CTATCC	hexa	118,247 -> 118,259	6-12	<i>rps12-trnV</i>	(CTATCC) <sub>1</sub>	N	(CTATCC) <sub>1</sub>	(CTATCC) <sub>2</sub>
TAACGA	hexa	119,644 -> 119,655	0-12	<i>rps12-trnV</i>	(TAACGA) <sub>0</sub>	(TAACGA) <sub>1</sub>	(TAACGA) <sub>2</sub>	(TAACGA) <sub>1</sub>
AATATT	hexa	170,676 -> 170,687	6-12	<i>rpl32</i> CDS	(AATATT) <sub>1</sub>	(AATATT) <sub>2</sub>	N	(AATATT) <sub>2</sub>
AAAAAAT	hepta	33,216 -> 33,235	7-14	<i>rpoB-trnC</i>	(AAAAAAT) <sub>1</sub>	(AAAAAAT) <sub>1</sub>	(AAAAAAT) <sub>2</sub>	N
AGATCAT	hepta	46,453 -> 46,468	7-14	<i>trnM-rps14</i>	(AGATCAT) <sub>1</sub>	(AGATCAT) <sub>2</sub>	N	(AGATCAT) <sub>1</sub>
TTTTCTTCT	nona	147,707 -> 147,725	9-18	<i>rps15</i> CDS	(TTTTCTTCT) <sub>2</sub>	(TTTTCTTCT) <sub>2</sub>	(TTTTCTTCT) <sub>1</sub>	(TTTTCTTCT) <sub>2</sub>
AAATGTTCC	nona	176,491 -> 176,508	9-18	<i>ycf2</i> CDS	(AAATGTTCC) <sub>1</sub>	(AAATGTTCC) <sub>2</sub>	N	(AAATGTTCC) <sub>1</sub>

De acordo com a literatura disponível sobre marcadores genéticos populacionais, vários SSRs já foram descritos para orquídeas sapatinhos-de-vênus, como 24 *loci* de microssatélites polimórficos em *Paphiopedilum rothschildianum* (RODRIGUES e KUMAR, 2009), 10 *loci* de microssatélites polimórficos em *Paphiopedilum concolor* (LI et al., 2010) e 15 *loci* de microssatélites para *Cypripedium calceolus* (MINASIEWICZ e ZNANIECKA, 2014). Os microssatélites polimórficos descritos nos estudos de Minasiewicz e Znaniecka (2014), Li et al. (2010) e Rodrigues e Kumar (2009) estavam localizados em regiões aleatórias do DNA. Os métodos de detecção de SSRs utilizados por estes autores não informam se os mesmos se encontravam em regiões gênicas ou intergênicas do cpDNA, nDNA ou mtDNA. Para a caracterização e distinção de populações esta informação não é de extrema importância, apesar de ser um dado muito interessante e importante para outros tipos de estudos. Utilizando técnicas de amplificação e sequenciamento de SSRs flanqueados por regiões conservadas, Fay et al. (2009) obtiveram sucesso em distinguir 23 indivíduos de *Cypripedium calceolus*. Os autores encontraram microssatélites, *indels* e Polimorfismos de Nucleotídeo Único (*Single Nucleotide Polymorphisms* - SNPs) em regiões plastidiais não-codantes, como *accD-psa1* e *rps16 intron*.

De todas as regiões polimórficas encontradas para *Paphiopedilum barbatum*, SSRs com *motifs* de dinucleotídeos e hexanucleotídeos parecem ser mais promissoras para posterior amplificação de um conjunto maior de amostras de *P. barbatum* de diferentes regiões. SSRs com *motifs* de dinucleotídeos mostraram ser eficientes para estudos de genética de população de orquídeas sapatinhos-de-vênus. SSRs com *motifs* de hexanucleotídeos foram os mais frequentes entre espécimes de *Paphiopedilum barbatum*. Portanto, *primers* devem ser desenvolvidos nas regiões flangeadoras das sequências envolvidas, em particular: *rps12-trnV*, *trnR-rrn5*, *ndhB-rps7*, *psbC-atpB*, *rbcL*, *ycf3*, *psbC*, *rpoB*, *rps12(intron)* e *trnI* (GAU) (*intron*). SSRs localizadas nestas regiões estão grifadas em cinza na Tabela 9.

## 5.5. MÉRITOS SOBRE A UTILIZAÇÃO DE DIFERENTES ESTRATÉGIAS DE SEQUENCIAMENTO E MONTAGEM DE cpDNA

Ao comparar técnicas NGS com Sanger FGS, este último mostrou-se superior em termos de precisão. Há muitas limitações ao montar a sequência plastidial completa utilizando apenas dados de sequência NGS.

As técnicas de sequenciamento PCR de Longo Alcance (*Long Range PCR*), IonTorrent e Sanger são normalmente utilizadas de forma complementar para preencher lacunas e confirmar regiões com baixa cobertura (DOORDUIN et al., 2011). NGS, entretanto, permite a produção de altos volumes de dados de sequenciamento a baixos preços (METZKER, 2010), dando uma nova dimensão a estudos filogenéticos e de genética de populações. Mesmo quando não há sequenciamento adicional (*'shallow sequencing'*), o produto da montagem de dados NGS pode, ainda, ser valioso para estudos filogenéticos (ver *'genome skimming approaches'*) (STRAUB et al., 2012) e também para a rápida detecção de marcadores populacionais.

No presente trabalho, seqüências completas do plastoma não puderam ser obtidas, demonstrando ser uma limitação da técnica de sequenciamento NGS e também do alto nível de fragmentação de DNA de espécimes de herbário. Entretanto, os resultados mostraram que o processamento de dados de sequenciamento NGS, utilizando-se de diferentes estratégias de bioinformática, produziu resultados satisfatórios para realização de análise comparativa filogenética e populacional, permitindo a recuperação da maior parte dos genomas plastidiais sequenciados sem sequenciamento adicional.

Orchidaceae é uma família de Angiospermas rica em espécies, com mais de 30.000 espécies estimadas. Apesar de sua alta diversidade, microssatélites estão disponíveis somente para algumas espécies e são, na sua maioria, desenvolvidos usando métodos de sequenciamento Sanger (PANDEY e SHARMA, 2012). Pandey e Sharma (2012) usaram pela primeira vez, em orquídeas, sequenciamento NGS 454 Roche GS-FLX para isolar microssatélites em *Cypripedium kentuckiense* e *Pogonia ophioglossoides*. Dentre as tecnologias NGS, a plataforma 454 Roche GS-FLX gera *reads* relativamente mais longas do que outras técnicas, sendo então mais eficiente para detecção de microssatélites. Ainda assim, se uma região repetitiva for maior que o tamanho de leitura de 454 Roche NGS (*reads* costumam ser de aproximadamente 400pb), seu tamanho não poderá ser inferido com precisão e será então subestimado (PANDEY e SHARMA, 2012), além de ser muito complexo projetar *primers* para tal seqüência. Estes autores encontraram 20,697 SSRs, sendo a maioria adequada para posterior desenho de *primers* (79% das SSRs), demonstrando que métodos NGS pode também modificar dimensões em estudos de genética de populações.

A combinação de métodos de sequenciamento e de bioinformática utilizada no presente trabalho fornece novo escopo para a rápida detecção de novos marcadores moleculares a baixos custos e demonstrou ser

efetiva para espécimes de herbário e material fresco. O presente estudo conseguiu detectar com sucesso 29 marcadores putativos para estudos filogenéticos em Cyripedioideae e 54 SSRs potenciais marcadores promissores para estudos populacionais de *Paphiopedilum barbatum*. Como continuação do trabalho, *primers* devem ser projetados nas regiões flanqueadoras das sequências envolvidas. Os marcadores filogenéticos aqui sugeridos devem ser utilizados para amplificação de DNA de espécimes de todos os cinco gêneros de Cyripedioideae, com a inclusão dos gêneros *Meximepium* e *Selenipedium*, incluindo o máximo de terminais em um alinhamento. Após a produção deste, análise filogenética utilizando análise Bayesiana e *Maximum Likelihood* deve ser realizada.

No caso de marcadores moleculares para genética de populações, *primers* devem ser projetados nas regiões flanqueadoras dos microssatélites aqui sugeridos. Além disto, um maior número de espécimes deve ser incluído no alinhamento. Como espécimes de herbário também podem ser utilizados para análise comparativa, a amostragem poderia ser, teoricamente, expandida para mais 18 indivíduos depositados em coleções no mundo todo. Uma pesquisa rápida sobre espécimes de herbário de *Paphiopedilum barbatum* disponíveis permitiu a detecção de 18 indivíduos de localidades diferentes que poderiam ser inclusos em estudos futuros (Tabela 10). Os espécimes encontrados estão descritos na Tabela 10 e ilustrados nas Figuras 30, 31 e 32.

## 5.6. VALOR ACRESCENTADO PELA UTILIZAÇÃO DE ESPÉCIMES DE HERBÁRIO

Herbários representam grandes depositórios de DNA que podem ser explorados, tanto para a obtenção de genes quanto para obtenção de genomas inteiros (BAKKER et al., 2015). Esforços para obter informação genética de espécimes de herbário são muitas vezes prejudicados pelos altos níveis de degradação característicos de DNA histórico, impossibilitando a obtenção de fragmentos amplificáveis (STAATS et al., 2011). Técnicas de NGS utilizam moléculas curtas de 100-300pb, sendo favorável para o sequenciamento de DNA degradado (STAATS et al., 2013). Dados de sequenciamento de qualidade podem ser obtidos a partir de espécimes de herbário, proporcionando oportunidades de explorar DNA histórico em contexto filogenético. Bakker et al. (2015) processaram dados de sequência *Illumina* de 93 espécimes (73 de herbário), pertencentes a 12 famílias de Angiospermas, e produziram *assemblies* para 74 espécimes. Os autores demonstraram que as

sequências produzidas a partir de DNA histórico tinham menores valores de cobertura e eram um pouco menos alongadas, mas, ainda assim, eram comparáveis com aquelas geradas a partir de DNA moderno.

Os dados de sequência dos dois espécimes de herbário de *P. barbatum*, utilizados no presente trabalho, foram processados e foi possível recuperar 59,52% e 21,50% do cpDNA de cada amostra. Foram recuperados 93,45% do cpDNA da amostra fresca *P. barbatum* e, portanto, foi encontrada uma diferença significativa entre a qualidade de sequenciamento e montagem de dados NGS entre DNA atual e histórico. Estudos anteriores não encontraram diferenças significativas no tamanho de *assemblies* de amostras frescas e de herbário (STAATS et al., 2013; BAKKER et al., 2015; LEI et al., 2015).

Desta forma, apesar do número restrito de cpDNA *reads* produzido a partir do sequenciamento de DNA histórico, dados de sequência confiáveis e informativos puderam ser produzidos. Neste estudo, foi possível produzir *assemblies* que, mesmo fragmentadas e ainda repletas de lacunas, proporcionaram dados suficientes para realização de análise comparativa a nível de subfamília e mesmo a nível populacional. Marcadores filogenéticos e populacionais puderam ser detectados de maneira efetiva. Para dar continuidade nos estudos populacionais de *Paphiopedilum barbatum*, propõe-se a inclusão de sequências SSRs, detectadas neste trabalho, incluindo um maior número de indivíduos. Como muitas outras espécies do gênero, *Paphiopedilum barbatum* está no *Orchid Checklist* de espécies ameaçadas de extinção. Neste caso, o sequenciamento de amostras de herbário torna-se ainda mais valioso.

**Tabela 10:** Espécimes de *Paphiopedilum barbatum* pertencentes a coleções *in vivo* (cultivadas) e de herbário (excicatas ou *spirit*) disponíveis em diferentes inventários e provenientes de localidades diversas.

Espécime	Invetário	Estado	Coletor	nº do coletor	Ano de coleta	País	Localidade
K000363770	Royal Botanical Gardens Kew (RBG – Kew)	excicata	Griffth	s.n.	s.d.	Malásia Peninsular	Mount Ophir
K000363771	(RBG – Kew)	excicata	Griffth	s.n.	s.d.	Malásia Peninsular	Mount Ophir
K000363769	(RBG – Kew)	excicata	Cuming	s.n.	s.d.	Malásia	Malacca
W339909	(RBG – Kew)	excicata	Anônimo	s.n.	Mar-1976	Indonésia	Borneo
W0045346	Herbarium WU, University of Vienna	excicata	Anônimo	s.n.	Mar- 1976	Indonésia	Borneo
RB00258494	Jardim Botânico do Rio de Janeiro	excicata	Mohd Shah	MS.3292	Jun-1974	Malásia	Malacca
P00333272	Muséum National d'Histoire Naturelle (MNHN)	excicata e <i>in vivo</i>	C.d'Alleizette	s.n.	1838	Malásia	Kao Katakuram Rang-nga Khao Pawta Suang Kaew, Panaung Kao kao Songkla Pulau Pinang
P00333275	MNHN	excicata	A.F.G. Kerr	797	Ago-1930	Tailândia	
P00333274	MNHN	excicata	A.F.G. Kerr	701	s.d.	Tailândia	
P00333273	MNHN	excicata	A.F.G. Kerr	624	Jul-1928	Tailândia	
6052	Smithsonian National Museum of Natural history (SNMNH)	excicata ou <i>spirit</i>	Sidek Bin Kiah	S.230	Abr-1968	Malásia	District of Columbia
227863	SNMNH	<i>in vivo</i>	Pfister, H.	s.n.	1887	Estados Unidos	District of Columbia
227862	SNMNH	<i>in vivo</i>	Harrison, C. Larsen, K.	s.n.	Abr-1894	Estados Unidos	Phangnga
6051	SNMNH	<i>in vivo</i>	Larsen, S. S.	33472	Mar-1974	Tailândia	Kedah park
706010	National Herbarium of New South Wales (NSW)	<i>in vivo</i>	P. Vaughan	s.n.	Ago-1981	Malásia	Kedah park
206680	NSW	<i>in vivo</i>	P. Vaughan	s.n.	Ago-1981	Malásia	Phangnga
2755	Herbarium Jany Renz	excicata e <i>in vivo</i>	Anônimo	s.n.	Mar-1945	Malaysia	unknown



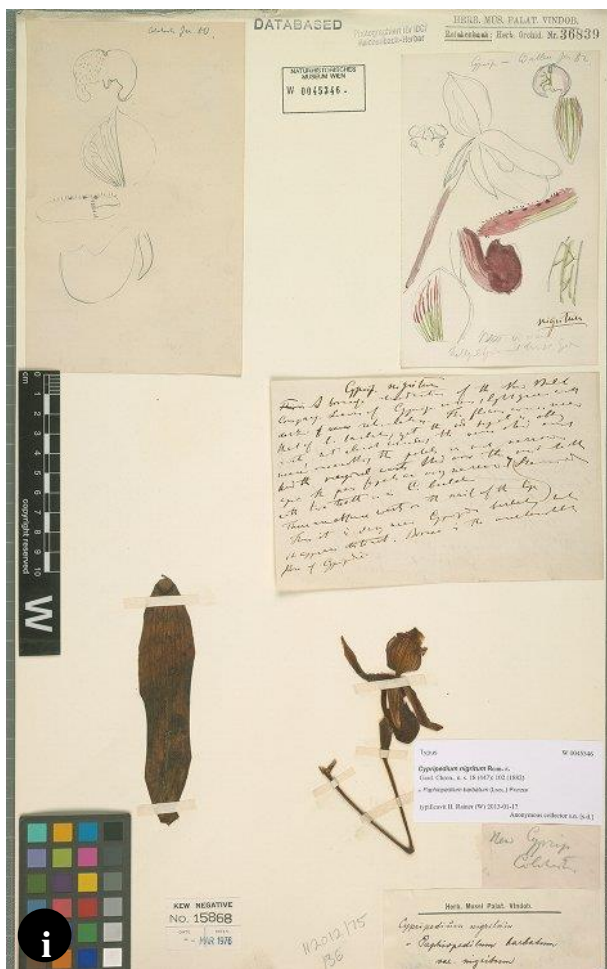
**Figura 30- a-d.** Espécimes disponíveis de *Paphiopedilum barbatum*. **a.** Espécime 2755, exsiccata. Coletado na Malásia (localidade desconhecida) em março de 1945. O espécime também encontra-se cultivado. Coleções *in vivo* e de herbário da fundação suíça para orquídeas *Swiss Orchid Foundation at the Herbarium Jany Renz*. **b-c.** Coletados na Malásia Peninsular, Mont Ophir (data desconhecida). Inventário do herbário do Jardim Botânico de Kew (*Royal Botanic Gardens, Kew*). **b.** Espécime K00363770. **c.** Espécime K00363771. **d.** Espécime RB00258494, coletado em junho de 1974, na Malásica, região de Malacca. Inventário do herbário do Jardim Botânico do Rio de Janeiro.



**Figura 31- e-f.** Espécimes disponíveis de *Paphiopedilum barbatum*. **e-h.** *Muséum National d'histoire naturelle*, Paris. **e.** Espécime P00333272, coletado na Malásia, região de Kao Katakuram Rangnga, em 1838. **f.** Espécime P00333274, coletado na Tailândia, região de Kao Kao Songkla (data desconhecida). **g.** Espécime P00333273, coletado na Tailândia, região de Pulao Pinang, em julho de 1928. **h.** Espécime P00333275, coletado na Tailândia, região de Khao Pawta Suang Kaew Panaung, em agosto de 1930.



**Figura 32 - i.** Espécime W0045346, *Paphiopedilum barbatum*, coletado na Indonésia, região de Boreno, em Março de 1976. Coleção de herbário da universidade de Viena (*Herbarium WU, University of Vienna*).



## 6. CONCLUSÕES

A combinação de métodos de sequenciamento e de bioinformática utilizada no presente trabalho fornece novo escopo para a rápida detecção de novos marcadores moleculares a baixos custos, demonstrando ser efetiva para espécimes de herbário e material fresco. Considerando o número de regiões polimórficas encontradas, recomenda-se a obtenção de seqüências inteiras do genoma do cloroplasto para detectar diferenças genéticas dentro de uma espécie, bem como para inferir relações filogenéticas. Este trabalho detectou regiões putativas do cpDNA para posterior desenvolvimento de *primers*, podendo ser utilizadas em futuras investigações filogenéticas sobre a diversificação de orquídeas sapatinhos-de-vênus, e também em futuras investigações sobre estruturas populacionais de *Paphiopedilum barbatum*.

## REFERÊNCIAS

ALBERT, V. A. Cladistic relationships of the slipper orchids (Cypripedioideae, Orchidaceae) from congruent morphological and molecular data. **Lindleyana**, v. 9, p. 115-132, 1994.

ÁLVAREZ, I.; WENDEL, J. F. Ribosomal ITS sequences and plant phylogenetic inference. **Molecular phylogenetics and evolution**, v. 29, p. 417-434, 2003.

BAKKER, F. T.; CULHAM, A.; DAUGHERTY, L. C.; GIBBY, M. A trnL-F based phylogeny for species of Pelargonium (Geraniaceae) with small chromosomes. **Plant Systematics and Evolution**, v. 216, p. 309-324, 1999.

BAKKER, F. T.; LEI, D.; YU, J.; MOHAMMADIN, S.; WEI, Z.; VAN DE KERKE, S.; GRAVENDEEL, B.; NIEUWENHUIS, M.; STAATS, M.; ALQUEZAR-PLANAS, D. E.; HOLMER, R. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly (IOGA) pipeline. **Journal of the Linnean Society**, p. 1-11, 2015.

BALDAUF, S. L. Phylogeny for the faint of heart: a tutorial. **Trends in Genetics**, v. 19, n. 6, 2003.

BENDICH, A. J. Circular Chloroplast Chromosomes: The Grand Illusion. **The Plant Cell**, v. 16, p. 1661-1666, 2004.

BENJAMINI, Y.; SPEED, T. P. Summarizing and correcting CG content bias in high-throughput sequencing. **Nucleic Acids Research**, p. 1-14, 2012.

CAMERON, K. M. A comparison of plastid atpB and rbcL gene sequences for inferring phylogenetic relationships within Orchidaceae. In: J. T., E. A. F. J. M. P. L. M. P. A. M. G. S. **Monocots: a comparative biology and evolution**. [S.l.]: [s.n.], v. 2, 2006. p. 447-464.

CAMERON, K. M. Molecular phylogenetics of Orchidaceae: the first decade of DNA sequencing. **Orchid Biology Reviews and Perspectives**, Bronx, NY, v. XI, p. 163-200, 2007.

CAMERON, K. M. CHASE, M. W.; WHITTEN, W. M.; KORES, P. J.; JARRELL, D. C.; ALBERT, V. A.; YUKAWA, T.; HILLS, H. G.; GOLDMAN, D. H. A phylogenetic analysis of the Orchidaceae: evidence from *rbcL* nucleotide sequences. **American Journal of Botany**, v. 86, n. 2, p. 208-224, 1999.

CHANG, C-C.; LIA, H-C.; LIN, I-P.; CHOW, T-Y.; CHEN, H-H.; CHEN, W-H.; CHENG, C-H.; LIN, C-Y.; LIU, S-M.; CHANG, C-C.; CHAW, S-M. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative Analysis of Evolutionary Rate with that of Grasses and Its Phylogenetic Implications. **Molecular Biology and Evolution**, v. 23(2), p. 279-291, 2006.

CHASE, M. W.; SOLTIS, D. E.; OLMSTEAD, R. G.; MORGAN, D.; LES, D. H.; MISHLER, B. D.; DUVALL, M. R.; PRICE, R. A.; HILLS, H. G.; QIU, Y-L.; KRON, K. A.; RETTIG, J. H.; CONTI, E.; PALMER, J. D.; MANHART, J. R.; SYTSMA, K. J.; MICHAELS, H. J.; KRESS, J. W.; KAROL, K. G.; CLARK, D. W.; HEDREN, M.; GAUT, B. S.; JANSEN, R. K.; KIM, K-J.; WIMPEE, C. F.; SMITH, J. F.; FURNIER, G. R.; STRAUSS, S. H.; XIANG, Q-Y.; PLUNKETT, G. M.; SOLTIS, P. S.; SWENSEN, S. M.; WILLIAMS, S. E.; GADEK, P. A.; QUINN, C. J.; EGUIARTE, L. E.; GOLENBERG, E.; LEARN, J. R.; GERALD, H.; GRAHAM, S. W.; BARRETT, S. C. H.; DAYANANDAN, S.; ALBERT, V. A. Phylogenetics of Seed Plants: An analysis of Nucleotide Sequences from the Plastid Gene *rbcL*. **Annals of the Missouri Botanical Garden**, v. 80, n. 3, p. 528-548+550-580, 1993.

CHASE, M. W. CAMERON, K. M.; HILLS, H. G.; JARRELL, D. **Molecular systematics of the Orchidaceae and other lilioid monocots**. Proceedings of the 14th World Orchid Conference. London: HMSO: Pridgeon A. ed. 1994. p. 61-73.

CHASE, M. W.; CAMERON, K. M.; BARRETT, R. L.; FREUDENSTEIN, J. V. DNA data and Orchidaceae systematics: a new phylogenetic classification. In: DIXON, K. W., K. S. B. R. C. P. **Orchid Conservation**. Borneo: Kota Kinabalu: Natural History Publications, 2003. p. 69-89.

CHASE, M. W.; WILLIAMS, N. H.; DE FARIA, A. D.; NEUBIG, K. M.; AMARAL, M. C. E.; WHITTEN, M. W. Floral convergence in

Oncidiinae (Cymbidieae; Orchidaceae): an expanded concept of Gomesa and a new genus Nohawilliamsia. **Annals of Botany**, v. 104, n. 3, p. 387-402, 2009.

CHOCHAI, A.; LEITCH, I. J.; INGROUILLE, M. J.; FAY, M. Molecular phylogenetics of Paphiopedilum (Cypripedioideae; Orchidaceae) based on nuclear ribosomal ITS and plastid sequences. **Botanical Journal of the Linnean Society**, v. 170, p. 176-196, 2012.

CLARK, S. C.; EGAN, R.; FRANZIER, P. I.; WANG, Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. **Bioinformatics**, v. 29, n. 4, p. 435-443, 2013.

COX, A. V.; PRIDGEON, A. M.; ALBERT, V. A.; CHASE, M. W. Phylogenetics of the slipper orchids (Cypripedioideae, Orchidaceae): nuclear rDNA ITS sequences. **Plant Systematics and Evolution**, v. 208, p. 197-223, 1997.

CRIBB, P. **The genus Cypripedium**. Portland: Timber Press, 1997. 294 p.

DAYARIAN, A.; MICHAEL, T. P.; SENGUPTA, A. M. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. **BioMed Central Bioinformatics**, v. 11, n. 345, p. 1-21, 2010. Disponível em: <<http://www.biomedcentral.com/1471-2105/11/345>>.

DELANNOY, E.; FUJII, S.; DES FRANCS-SMALL, C.; BRUNDRETT, M.; SMALL, I. Rampant Gene Loss in the Underground Orchid Rhizantella Gardeni Highlights Evolutionary Constraints on Plastid Genomes. **Molecular Biology and Evolution**, v. 28, n. 7, p. 2077-2986, 2011.

DOORDUIN, L.; GRAVENDEEL, B.; LAMMERS, Y.; ARIYREK, Y.; CHIN-A-WOENG, T.; VRIELING, K. The complete chloroplast genome of 17 individuals of pest species Jacobea vulgaris: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. **DNA RESEARCH**, 18, n. 2, 2011. 93-105.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v. 19, p. 11-15, 1987.

DRESSLER, R. L. **Phylogeny and the Classification of the Orchid Family**. [S.l.]: Cambridge University Press, 1993.

FATIHAH, N. H.; FAY, M. F.; MAXTED, N. Molecular Phylogenetics of *Cypripedium* L. (Cypripedioideae: Orchidaceae) Based on Plastid and Nuclear DNA Sequences. **Journal of Biotechnology**, v. 2, p. 35-51, 2011.

FAY, M. F.; BONE, R.; COOK, P.; KAHANDAWALA, I.; GREENSMITH, J.; HARRIS, S.; PEDERSEN, H. E.; INGROUILLE, M. J.; LEXER, C. Genetic diversity in *Cypripedium calceolus* (Orchidaceae) with a focus on north-western Europe, as revealed by plasmid DNA length polymorphisms. **Annals of Botany**, v. 104, p. 517-525, 2009.

FAY, M. F.; CHASE, W. M. Orchid biology: from Linnaeus via Darwin to the 21st century. **Annals of Botany**, 104, 2009. 359–364.

FRANCKLIN, R.; GOSLING, R. Molecular Configuration in Sodium Thymonucleate. **Nature**, v. 171, p. 740-741, 1953.

FREUDENSTEIN, J. V.; VAN DEN BERG, C.; GOLDMAN, D. H.; KORES, P. J.; MOLVRAY, M.; CHASE, M. W. An expanded plastid DNA phylogeny of Orchidaceae and analysis of jackknife branch support strategy. **American Journal of Botany**, v. 91, n. 1, p. 149-157, 2004.

GHOORBANI, A.; GRAVENDEEL, B.; NAGHIBI, F.; DE BOER, H. Wild orchid tuber collection in Iran: a wake-up call for conservation. **Biodiversity and conservation**, v. 23, p. 2749-2760, 2014.

GRAVENDEEL, B.; SMITHSON, A.; SLIK, F. T. W.; SCHUITEMAN. Epiphytism and Pollinator Specialization: Drivers for Orchid Diversity? **Philosophical Transactions: Biological Sciences**, v. 359, n. 1450, p. 1523-1535, 2004.

GUO, Y-Y.; LUO, Y-B.; LIU, Z-J.; WANG, X-Q. Evolution and Biogeography of the Slipper Orchids: Eocene Vicariance of the



Conduplicate Genera in the Old and New World Tropics. **PLoS ONE**, 7(6), 2012. e:38788.

HAHN, C.; BACHMANN, L.; CHEVREUX, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. **Nucleic Acids Research**, v. 41, n. 13, p. 1-9, 2013. Disponível em: <<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3711436/pdf/gkt371.pdf>>.

HASTON, E.; RICHARDSON, J. E.; STEVENS, P. F.; CHASE, M. W.; HARRIS, D. J. The Linear Angiosperm Phylogeny Group (LAPG) III: a linear sequence of the families in APG III. **Botanical Journal of the Linnean Society**, v. 161, p. 128-131, 2009.

HOLMER, R.; NIEUWENHUIS, M. Iterative Organelle Genome Assembly. **Unpublished**, 2014.

HUNT, M.; NEWBOLD, C.; BERRIMAN, M.; OTTO, T. D. A comprehensive evaluation of assembly scaffolding tools. **Genome Biology**, v. 15, n. R42, 2014. Disponível em: <<http://genomebiology.com/2014/15/3/R42>>.

HUTCHISON III, C. A. DNA sequencing: bench to bedside and beyond. **Nucleic Acids Research**, v. 35, n. 18, p. 6227-6237, 2007.

JHENG, C-F.; CHEN, T-C.; LIN, J-Y.; CHEN, T-C.; WU, W-L.; CHANG, C-C. The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish Phalaenopsis orchids. **Plant Science**, v. 190, p. 62-73, 2012.

JUDD, W. S.; CAMPBELL, C. S.; KELLOG, E. A.; STEVENS, P. F.; DONOGHUE, M. J. **Plants Systematics: A phylogenetic approach**. 2. ed. Sunderland, Massachusetts: Sinauer, 2002.

KING, H. geology.com. Acesso em: 23 maio 2015.

LI, L-N.; ZENG, S.; ZHENG, F.; CHEN, Z-L.; WU, K-L.; ZHANG, J-X.; DUAN, J. Isolation and Characterization of 10 Polymorphic Microsatellite Loci in *Paphiopedilum concolor* (Batem.) Pfitzer

(Orchidaceae) and Cross-species Amplification. **HortScience**, v. 45, n. 8, p. 1286-1287, 2010.

LIN, C.-S.; CHEN, J.W.; HUANG, Y-T.; CHAN, M-T.; DANIELL, H.; CHANG, W-J.; HSU, C-T.; LIAO, D-C.; WU, F-H.; LIN, S-Y.; LIAO, C-F.; DEYHOLOS, M. K.; WONG, G. K-S.; ALBERT, V. A.; CHOU, M-L.; CHEN, C-Y.; SHIH, M.-C. The location and translocation of *ndh* genes of chloroplast origin in the orchidaceae family. **Nature Scientific Reports**, v. 5, n. 9040, p. 1-10, 2015.

MÄKINEN, V.; SALMELA, L.; YLINEN, J. Normalized N50 assembly metric using gap-restricted co-linear chaining. **BioMed Central Bioinformatics**, v. 13, n. 255, p. 1-5, 2012.

MANDRIC, I.; ZELIKOVSKY, A. ScaffMatch: scaffolding algorithm based on maximum weight matching. **Bioinformatics**, 2015.

MAXAM, A. M.; GILBERT, W. A new method for sequencing DNA. **Proceedings of the National Academy of Sciences USA**, v. 74, p. 560-564, 1977.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Review**, v. 11, p. 31-46, 2010.

MINASIEWICZ, J.; ZNANIECKA, J. M. Characterization of 15 novel microsatellite loci for *Cypripedium calceolus* (Orchidaceae) using MiSeq sequencing. **Conservation Genetics Resources**, v. 6, p. 527-529, 2014.

NEUBIG, K. M.; WHITTEN, W. M.; WILLIAN, N. H.; BLANCO, M. A.; ENDARA, L.; BURLEIGH, J. G.; SILVEIRA, K.; CUSHMAN, J. C.; CHASE, M. W. Generic recircumscriptions of Oncidiinae (Orchidaceae: Cymbidieae) based on maximum likelihood analysis of combined DNA datasets. **Botanical Journal of the Linnean Society**, v. 168, p. 117-146, 2012.

PAKENDORF, B.; NOVGORODOV, I. N.; OSAKOVSKIJ, V. L.; DANILOVA, A. P.; PROTOD'JAKONOV, A. P.; STONEKING, MARK. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. **Human Genetics**, v. 120, p. 334-353, 2006.

PAN, I-C.; LIAO, D-C.; WU, F-H.; DANIELL, H.; SINGH, N.D.; CHANG, C.; SHIH, M-C.; CHAN, M-TI-C.; LIN, C-S. Complete Chloroplast Genome Sequence of an Orchid Model Candidate: *Erycina pusilla* Apply in Tropical *Oncidium* Breeding. **PLoS ONE**, v. 7(4), p. 1-12, 2012.

PANDEY, M.; SHARMA, J. Efficiency of microsatellite isolation from orchids via next generation sequencing. **Open Journal of Genetics**, v. 2, p. 167-172, 2012.

PFITZER, E. H. Orchidaceae-Pleonandrae. In: ENGLER, A. **Das Pflanzenreich**. [S.l.]: Leipzig:Engelmann, 1903. p. 1-132.

PRIDGEON, A. M.; CRIBB, P. J.; CHASE, M. W.; RASMUSSEN, F. N. **Genera Orchidacearum**. [S.l.]: Oxford University Press, v. 5 Epidendroideae (part II), 2009.

PRIDGEON, A. M.; CRIBB, P. J.; CHASE, M. W.; RASMUSSEN, F. N. **Genera Orchidacearum**. [S.l.]: Oxford University Press, v. 1 General introduction, Apostasioideae, Cyripedioideae, 1999.

PRIDGEON, A. M.; CRIBB, P. J.; CHASE, M. W.; RASMUSSEN, F. N. **Genera Orchidacearum**. [S.l.]: Oxford University Press, v. 2 Orchidoideae (part 1), 2001.

PRIDGEON, A. M.; CRIBB, P. J.; CHASE, M. W.; RASMUSSEN, F. N. **Genera Orchidacearum**. [S.l.]: Oxford University Press, v. 3 Orchidoideae (part II) and Vanilloideae, 2003.

PRIDGEON, A. M.; CRIBB, P. J.; CHASE, M. W.; RASMUSSEN, F. N. **Genera Orchidacearum**. [S.l.]: Oxford University Press, v. 4 Epidendroideae (part I), 2005.

PRIDGEON, A. M.; CRIBB, P. J.; CHASE, M. W.; RASMUSSEN, F. N. **Genera Orchidacearum**. [S.l.]: Oxford University Press, v. 6 Epidendroideae (Part III), 2014.

RODRIGUES, K. F.; KUMAR, S. V. Isolation and characterization of 24 microsatellite loci in *Paphiopedilum rothschildianum*, an endangered slipper orchid. **Conservation Genetics**, v. 10, p. 127-130, 2009.

RUHLMAN, T. A.; CHANG, W-J.; CHEN, J. J. W.; HUANG, Y-T.; CHAN, M-T.; ZHANG, J.; LIAO, D-C.; BLAZIER, J. C.; XIAOHUA, J.; SHIH, M-C.; JANSEN, R. K.; LIN, C-S. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. **BioMed Central Plant Biology**, v. 15 (100), p. 1-9, 2015.

SAGAN, L. On the Origin of Mitosing Cells. **Journal of Theoretical Biology**, v. 14, 1967.

SAHLIN, K.; STREET, N.; LUNDEBERG, J.; ARVESTAD, L. Improved gap size estimation for scaffolding algorithms. **Bioinformatics**, v. 28, n. 17, p. 2215-2222, 2012.

SAHLIN, K.; VEZZI, F.; NYSTEDT, B.; LUNDEBERG, J.; ARVESTAD, L. BESST - Efficient scaffolding of large fragmented assemblies. **BioMed Central Bioinformatics**, v. 15, n. 281, p. 1-11, 2014. Disponível em: <<http://www.biomedcentral.com/1471-2105/15/281>>.

SANGER, F.; NICKLEN, S.; COULSON, A. R. Nucleotide sequence of bacteriophage phiX174 DNA. **Nature**, v. 265, p. 687-695, 1977.

SANGER, F.; AIR, G. M.; BARREL, B. G.; BROWN, N. L.; COULSON, A. R.; FIDDES, J. C.; HUTCHISON, C. A.; SLOCOMBE, P. M.; SMITH, M. Nucleotide sequence of bacteriophage lambda DNA. **Journal of Molecular Biology**, v. 162, p. 729-773, 1982.

SANGER, F.; COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of Molecular Biology**, v. 25;94(3), p. 441-448, 1975.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences USA**, v. 74, p. 5463-5467, 1977.

SCHUSTER, S. C. Next-generation sequencing transforms today's biology. **Nature Methods**, v. 5, n. 1, 2008. Acesso em: 18 maio 2015.

SEEMAN, T. **De novo genome assembly of NGS data**. Victorian Bioinformatics Consortium. [S.l.]: Monash University. 2011. p. 1-45.

SHAWN, J.; LICKEY, E.B.; SCHILLING, E.E.; SMALL, R.L. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. **American Journal of Botany**, v. 94, n. 3, p. 275-288, 2007.

SIMPSON, M. G. **Plant Systematics**. 2nd. ed. [S.l.]: Academic Press, 2010.

SINGER, R. B.; GRAVENDEEL, B.; CROSS, H.; RAMIREZ, S. R. The use of orchid pollinia or pollinaria for taxonomic identification. **Selbyana**, v. 29, n. 1, p. 6-19, 2008.

SINSHEIMER, R. L. A single-stranded DNA from bacteriophage phi X174. **Journal of MOlecular Biology**, v. 1, p. 43, 1959.

SOLTIS, D.E.; SMITH, S.A.; CELLINESE, N.; WURDACK, K.J.; TANK, D.C.; BROCKINGTON, S.F.; REFULIO-RODRIGUEZ, N.F.; WALKER, J.B.; MOORE, M.J.; CARLSWARD, B.S.; BELL, C.D.; LATVIS, M.; CRAWLEY, S.; BLACK, C.; DIOUF, D.; XI, D.; RUSHWORTH, C.A.; GITZENDANNER, M.A.; SYTSMA, K.J.; QIU, Y.L.; HILU, K.W.; DAVIS, C.C.; SANDERSON, M.J.; BEAMAN, R.S.; OLMSTEAD, R.G.; JUDD, W.S.; DONOGHUE, M.J.; SOLTIS, P.S. Angiosperm Phylogeny: 17 genes, 640 taxa. **American Journal of Botany**, v. 98(4), p. 704-730, 2011.

SOLTIS, D. E.; SOLTIS, P. S.; ZANIS, M. J. Phylogeny of seed plants based on evidence from eight genes. **American Journal of Botany**, v. 89, n. 10, p. 1670-1681, 2002.

SOLTIS, P. S.; SOLTIS, D. E.; CHASE, M. W. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. **Nature**, v. 402, p. 402-404, 1999.

STAATS, M.; CUENCA, A.; RICHARDSON, J.E.; GINKEL, R.V.; PETERSEN, G.; SEBERG; BAKKER, F. DNA Damage in Plant Herbarium Tissue. **PloS ONE**, v. 6(12), p. 1-9, 2011.

STAATS, M.; ERKENS, R.H.J.; VAN DE VOSSENBERG, B.; WIERINGA, J.J.; KRAAIJEVELD, K. Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. **PloS ONE**, v. 8, n. 7, p. 1-11, 2013.

STERN, W. L. Orchidaceae. In: GREGORY, M.; CUTLER, D. F. **Anatomy of the monocotyledons**. [S.l.]: Oxford University Press, v. 10, 2014.

STEVENS, P. F. [www.mobot.org/MOBOT/research/APweb/](http://www.mobot.org/MOBOT/research/APweb/). **Angiosperm Phylogeny Website version 12**, 2001 onwards. Acesso em: 2015.

SUBEDI, A.; CHAUDHARY, R. P.; VAN ACHTERBERG, C.; HEIJERMAN, T.; LENS, F.; VAN DOOREN, T. J. M.; GRAVENDEEL, B. Pollination and Protection against Herbivory of Nepalese Coelogyninae. **American Journal of Botany**, v. 98, n. 7, p. 1095-1103, 2011.

SUBEDI, A.; KUNWAR, B.; CHOI, Y.; DAI, Y.; VAN ANDEL, T.; CHAUDHARY, R. P.; DE BOER, H. J.; GRAVENDEEL, B. Collection and trade of wild-harvested orchids in Nepal. **Journal of Ethnobiology and Ethnomedicine**, v. 9, n. 64, p. 1-10, 2013.

VAN DEN BERG, C.; GOLDMAN, D. H.; FREUDENSTEIN, J. V.; PRIDGEON, A. M.; CAMERON, K. M.; CHASE, M. W. An overview of the phylogenetic relationships within Epidendroideae inferred from multiple DNA regions and recircumscription of Epidendreae and Arethuseae (Orchidaceae). **American Journal of Botany**, v. 92(4), p. 613-624, 2005.

VAN DEN BERG, C.; HIGGINGS, W. E.; DRESSLER, R. L.; WHITTEN, M.; SOTO-ARENAS, M. A. A phylogenetic study of Laeliinae (Orchidaceae) based on combined nuclear and plastid DNA sequences. **Annals of Botany**, v. 104, p. 417-430, 2009.

WATSON, J. D.; CRICK, F. H. Molecular structure of Nucleic Acids. **Nature**, v. 171, p. 737-738, 1953.

WILLIAMS, W. E.; GRIVET, C.; ZEIGER, E. Gas exchange in *Paphiopedilum*. **Plant Physiology**, v. 72, p. 906-908, 1983.

WU, F.-H.; CHAN, M.-T.; LIAO, D.-C.; HSU, C.-T.; LEE, Y.-W.; DANIELL, H.; DUVALL, M.R.; LIN, C.-S. Complete chloroplast genome of *Oncidium Gower Ramsey* and evaluation of molecular markers for identification and breeding in Oncidiinae. **BMC Plant**

**Biology**, v. 10, n. 68, p. 1-12, 2010. Disponivel em: <http://www.biomedcentral.com/1471-2229/10/68>.

YANG, J-B.; TANG, M.; LI, H-T.; ZHANG, Z-R.; LI, D-Z. Complete chloroplast genome of the genus *Cymbidium*: Lights into the species identification, phylogenetic implications and genetic analyses. **BioMedCentral Evolutionary Biology**, v. 13:84, p. 1-12, 2013. Disponivel em: <http://www.biomedcentral.com/1471-2148/13/84>.