Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular
Programa de Pós Graduação em Biologia Molecular

TESE DE DOUTORADO

Sequenciamento de DNA, montagem *de novo* do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de *Brachiaria ruziziensis*

Autor: Alexandre Magalhães Martins

Orientador: Dr. Márcio Elias Ferreira

Brasília, julho de 2013.

Alexandre Magalhães Martins

Sequenciamento de DNA, montagem *de novo* do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de *Brachiaria ruziziensis*

Tese apresentada ao Programa de Pós Graduação em Biologia Molecular da Universidade de Brasília como requisito para obtenção do Título de Doutor.

Orientador: Dr. Márcio Elias Ferreira

Brasília, julho de 2013.

Catalogação na fonte pela Biblioteca Universitária da
Universidade de Brasília

Dedico

À minha mãe Dona Edna Magalhães e ao meu pai Adalguacy Martins (*in memmoriam*).

AGRADECIMENTOS

À Universidade de Brasília pelo Instituto de Ciências Biológicas, aos professores e colaboradores do Programa de Pos-Graduação em Recursos Geneticos Vegetais, pela oportunidade e cooperação. Agradeço também aos professores do Departamento de Biologia Molecular da UNB, especialmente aos professores Dr. Marcelo Brígido e Dr. Renato de Oliveira Resende, aos colegas, e servidores.

À Embrapa Recursos Genéticos e Biotecnologia/Cenargen, pelo apoio no desenvolvimento dos trabalhos na pessoa grande amigo Dr. Márcio Elias Ferreira, que tanto se empenhou para que este trabalho fosse desenvolvido com excelência. Levarei comigo um grande exemplo de trajetória profissional, de trabalho em equipe, de profissionalismo e de respeito às pessoas.

Aos pesquisadores do Núcleo Temático de Recursos Genéticos da Embrapa Cenargen, que tive a oportunidade de conviver durante o curso, pela amizade e pelo suporte que de alguma forma muitos concederam, em especial, ao amigo Dr. Dário Grattapaglia, que gentilmente abriu as portas da Embrapa Cenargen para que eu pudesse realizar meu doutorado, acreditando na minha capacidade e pelo apoio que nunca me faltou. Ao Dr. Roberto Togawa, que sempre colocou o laboratório de bioinformática à disposição e que muito colaborou para os resultados. Aos pesquisadores Msc. Orzenil Júnior e ao Dr. Marco Pessoa Filho, que propuseram a colaborar em todos os momentos em que foram solicitados. A colega Msc. Ediene Gouveia que colaborou na validação dos marcadores desenvolvidos nesta tese e ao Msc. Pedro Tanno, com o qual trabalhei em parceria no primeiro capítulo. Agradeço aos profissionais, Dra. Vera Carneiro, Dr. Peter Inglis, que participaram disponibilizando material para pesquisa e análise.
À equipe de trabalho que auxiliou o desenvolvimento dos cruzamentos e fenotipagens das populações na pessoa do Dr. Paulo Hideo Nakano Rangel. À banca de defesa pelas sugestões e comentários, imprescindíveis para o enriquecimento deste trabalho. Ao Professor Robert Miller, ao professor Dr. Lúcio Flávio, à professora Dra. Maria Emilia e ao professor Dr. Paulo Hideo pela sua importante colaboração e enorme interesse pelo suporte na finalização do trabalho.

O meu sincero agradecimento à Bruna, Liamar, Rodrigo e Justino, por terem feito parte desta

etapa tão importante da minha vida. Gostaria de registrar aqui a convivência harmoniosa que tive com estes companheiros.

Quero agradecer finalmente à minha esposa Fernanda, aos meus filhos Thiago, Matheus, Isaac e Sara, a minha mãe Edna e minha irmã Leocilene, que abdicaram da minha companhia em muitos momentos, para que eu pudesse me dedicar aos estudos, sempre compreendendo e me apoiando neste intento. Por todas as oportunidades concedidas, fundamentais para o meu crescimento profissional e pessoal, serei eternamente grato.

Índice

I. LISTA DE TABELAS

Capítulo 1

Capítulo 2

assemblies of ruzigrass (*B. ruziziensis*) and rice (*Oryza sativa*) genomes, after classification of elements on different TE classes

Capítulo 3

1. Tabela 1- Métricas do sequenciamento e montagem do genoma de quatro espécies de Brachiaria usando *P. virgatum* (cp) como genoma de referência.

2. Tabela 2 – Parâmetros de sequenciamento e montagem de novo do genoma de quatro espécies de Brachiaria .

3. Tabela 3 – Número do scaffold e tamanho em número de bases da montagem de novo do cpDNA de quatro espécies de Brachiaria, que alinharam com o cpDNA de referência de P. virgatum (e-value = 0). Os scaffolds grifados em negrito correspondem as duas inverted repeats (IR) combinadas e foram consideradas em dobro para avaliação da cobertura linear.

4. Tabela 4 – Cobertura observada e tamanho (pb) de scaffolds obtidos na montagem de novo dos quatro cpDNA das espécies *B. ruziziensis, B. humidicola, B. brizantha e B. decumbens*. Os números que identificam os scaffolds correspondentes de cada espécie para as regiões IR, LSC e SSC do genoma do cloroplasto são apresentados.

5. Table 5 – Indel "primers" para Brachiaria desenvolvidos a partir da montagem de novo do genoma de cloroplasto para a identificação de espécies testados em gel de agarose. Os números de referência indicam a posição no genoma de cloroplasto de P. virgatum, números com dupla referência referem-se a posições em regiões IR. Marcadores entre as posições 107669 e 114885 estão em SSC após 81.616 estão na região LSC. *B. ruziziensis* (RUZI), *B. decumbens* (DEC), *B. brizantha* (BRI) e *B. humidicola* (HUM) e comprimento do fragmento esperado após amplificação.

6. Tabela 6 - Número de SSRs perfeitos com variações de di, tri e tetra nucleotídeos encontrados nas seqüências montadas de cpDNA de Brachiaria, tendo o genoma de cloroplasto de *P. virgatum* como referência. Os motivos mais abundantes são quantificados.

7. Tabela 7 - Número de indels e SNPs entre seqüências de cpDNA de quatro espécies de Brachiaria comparadas par-a-par. *B. humidicola* (Hum), *B. ruziziensis* (Ruzi), *B. decumbens* (Dec) e *B. brizantha* (Briz). A correlação entre Indel e SNPs é 0,856512.

8. Tabela 8. As estimativas de divergência evolutiva entre sequências completas de cpDNA de Brachiaria e outras gramíneas. As estimativas do erro padrão (s) são mostradas acima da diagonal. A análise envolveu oito seqüências de nucleotídeos. Todas as posições que contêm lacunas e dados faltantes foram eliminadas. Um total de 128.636 posições foi considerado no conjunto de dados final. Análises evolutivas foram realizadas usando MEGA5.

II. LISTA DE FIGURAS

Introdução

Capitulo 1

Capítulo 2

length of contig fraction > 200 bp, in Mpb; x axis = the number of contigs assembled / 1000).

3. Figure 3 – A database of 22,554 target *Brachiaria* sequences was used to query maize, sorghum and switch grass gene sequence databases. A total of 17,245 common gene orthologs identified between of *Brachiaria* and the three other grass species are depicted.

Capítulo 3

1. Figura 1 – Árvore filogenética obtida pelo método ML (Maximum Likelihood) após alinhamento de sequência de 741 bases da região nuclear 5.8S de rDNA (ITS) de seis acessos de Brachiaria representando quatro espécies (B. ruziziensis, B. brizantha, B. decumbens e B. humidicola). Log de verossimilhança = -1277.8197
2. Figura 2 - Árvore filogenética obtida pelo método ML (Maximum Likelihood) após alinhamento de sequência de 741 bases da região nuclear 5.8S de rDNA (ITS) de acessos de Brachiaria representando quatro espécies (*B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola*) e de acessos depositados no GenBank [3] (Gonzalez e Morthon, 2005). Filogenia inferida pelo modelo GTR (General Time Reversible model) (GTR). Log de verossimilhança = -2838.4822.
3. Figura 3 – Discriminação de acessos de quatro espécies de Brachiaria com marcadores indel selecionados no cpDNA. Polimorfismo de DNA de amostras de diferentes acessos do Banco de Germoplasma sobmetidas a eletroforese em gel de agarose 1%. Quatro marcadores são apresentados: 66584, 93252, 107669, além da combinação em multiplex dos marcadores 93252 e 107669. As amostras das diferentes espécies são apresentadas na seguinte ordem para cada marcador ou multiplex: *B. ruziziensis* (Kennedy, Colbase 2, Colbase 3*), B. brizantha* (Marandu, 591, 1384), *B. decumbens* (Basiliski, 116, 1058), *B. humidicola* (Tupi, 1929, 1937). Os marcadores são separados pela escada alélica (ladder) 50 pb (Promega).
4. Figura 4. Discriminação de acessos de quatro espécies de Brachiaria com marcador indel selecionado no DNAcp . Polimorfismo de DNA no loco indel RUBRIZ entre amostras de diferentes acessos do Banco de Germoplasma sobmetidas a eletroforese em gel de poliacrilamida. As amostras das diferentes espécies são apresentadas na seguinte ordem: *B. ruziziensis* (Kennedy, Colbase 2), *B. brizantha* (Marandu, 591), *B. decumbens* (Basiliski, 116), *B. humidicola* (Tupi, 1929). Os marcadores são

separados pela escada alélica (ladder) 50 pb (Promega). As amostras foram repetidas lado a lado, em testes de prova e contra-prova.

5. Figura 5. Mapa genético do genoma do cloroplasto de Brachiaria ruzizensis. O mapa inclui as repetições invertidas, IRa e IRb, regiões de cópia única pequena (SSC) e grande (LSC). Genes identificados no interior do mapa são transcritos no sentido horário, enquanto que os genes do exterior do mapa são transcritos em ordem inversa.
6. Figura 6. A história evolutiva foi inferida pelo método Maximum Likelihood baseado no modelo Tamura-Nei. As árvores com a maior verossimilhança (LSC =-18.409,3501, IR (combinado) = -12545.7330 e SSC = -28603.4495 são apresentadas. Árvore inicial (s) para a busca heurística foi obtida automaticamente através da aplicação de Neighbor-Join e algoritmos BioNJ a uma matriz de distâncias estimadas entre pares usando a abordagem de probabilidade de composição máxima (MCL) e, em seguida, selecionada a topologia com o valor de verossimilhança superior. A árvore está desenhada em escala, com comprimentos dos ramos medidos no número de substituições por sítio (acima dos braços). A análise envolveu quatro sequências de nucleotídeos. Posições do códon incluídos foram 1 2 ª 3 ª + não-codificante. Todas as posições que contêm lacunas e dados faltantes foram eliminados. Um total de 12.494, 8539 e 20350 posições foram consideradas no conjunto de final de dados.
7. Figura 7. A árvore filogenética construída por ML apresentando a maior verossimilhança = -243.395,9130. O modelo de variação da taxa foi evolutivamente invariável. Um total de 128.636 posições foram consideradas no conjunto de dados final. O relógio molecular foi calibrado usando um ponto de divergência de Oryza sativa e Zea mays com ocorrência há 65 milhões de anos (MYA). Taxa Evolutiva = 1,96807 X 10 -9. O número de repetições no teste de "bootstrap" foi 1000. Números entre parêntesis correspondem a estimativa de tempo de divergência com comprimentos medidos no número de substituições por sítio (acima dos ramos)
8. Figura 8 - A árvore filogenética construída por máxima parcimônia. A árvore mais parcimoniosa com comprimento = 11077 é apresentada. As percentagens de árvores idênticas computadas por teste de "bootstrap" (1000 réplicas) são mostrados ao lado dos ramos. Os comprimentos dos ramos foram calculados usando o método da média de percurso e está em unidades de número de mudanças ao longo de toda a sequência. Todas as posições que contêm lacunas e dados faltantes foram eliminadas da análise. Um total de 128.636 posições foi considerado no conjunto de dados final.

Sequenciamento de DNA, montagem *de novo* do genoma e desenvolvimento de marcadores microssatélites, indels e SNPs para uso em análise genética de *Brachiaria ruziziensis*

III. Introdução

O gênero Brachiaria

Braquiária é um gênero botânico pertencente à família Poaceae, subfamília Panicoidea, tribo Paniceae, descrito primeiramente por Trinius (1834) [4] como uma subdivisão de *Panicum*, posteriormente elevado à categoria de gênero por Grisebach (1853) [5]. Desde então, a classificação deste gênero tem sido: domínio Eukaryota, reino Plantae, superdivisão Spermatophyta, divisão Magnoliophyta, classe Liliopsida, subclasse Commelinidae, ordem Poales, família Poaceae e gênero *Brachiaria*.

O gênero *Brachiaria* caracteriza-se por possuir flor contendo de um a três estames, colmo herbáceo florescendo todos os anos, espiga unilateral ou panícula, espigueta comprimida dorsiventralmente, biflora, com o antécio terminal frutífero, o basal neutro ou masculino.

Diferenciação do gênero Brachiaria de outros gêneros de Poaceae

Embora o gênero tenha sido reconhecido no século XIX, a classificação botânica de braquiária não é considerada consistente em razão da dificuldade de definição clara de características morfológicas diferenciadoras. Os limites precisos para diferenciação de *Brachiaria* de gêneros próximos como *Urochloa*, *Eriochloa* e *Panicum* ainda geram dúvidas. Os principais caracteres que identificam o gênero *Brachiaria* dos outros gêneros próximos são as espiguetas de forma ovalada, arranjadas em racemos unilaterais, com a gluma inferior adjacente à ráquis. De acordo com Bogdan (1977) [6], as gramíneas do gênero *Brachiaria* "Signal" ou "Palisade grasses", são plantas perenes ou anuais, cespitosas ou decumbentes. A panícula consiste de poucos (às vezes um só) a diversos racemos com espiguetas sésseis ou subsésseis, arranjadas em duas fileiras em uma ráquis usualmente achatada. Dos dois flósculos da espigueta, o inferior é masculino com lema e pálea macios. O flósculo superior é fértil, bissexual ou muitas vezes feminino, achatado de um lado e convexo no outro. A cariopse está englobada dentro de um lema e pálea, duros e rígidos.

Segundo Rosengurtt et al. (1970) [1], o gênero *Brachiaria* apresenta panículas de espigas unilaterais de eixo alargado com espigueta mútica. A gluma I é adaxial. O Antécio II é coriáceo com asperezas punticuladas em finas linhas transversais. A panícula mede 11 a 24 cm, e contém de 3 a 7 espigas com espiguetas solitárias dispostas em duas fileiras. A ráquis

com 1,5 a 3 mm de largura possui pelos. A espigueta é obtusa de 4 a 4,6 mm. A gluma II e lema I, nervadas entre 5 e 8, sobrepassam quase 1 mm o antécio. A pálea II é neutra e a cariopse de 1,8 mm é pouco comprimida dorsiventralmente (Figura 1).

Nem sempre os taxonomistas concordam e, às vezes, essas características não são consistentes para todas as espécies do gênero e, por isso, aparecem os questionamentos. Por conseguinte, a taxonomia deste gênero não é satisfatória, tanto em relação à composição de suas espécies como na interrelação com outros gêneros [5].

Alcântara & Bufarah (1988) [7] e Soares Filho (1994) [8] apresentam a descrição morfológica das principais espécies cultivadas de *Brachiaria*. Porém, em ambos os trabalhos, a descrição é limitada, uma vez que se baseia em poucos acessos, não representando a variabilidade existente dentro de cada espécie [9].

Alguns genótipos têm sido amplamente distribuídos com o nome incorreto da espécie, criando confusão na literatura publicada [10]. Portanto, esses autores consideram que, à época, era preciso haver estudos morfológicos, agronômicos e moleculares detalhados para estabelecer a identidade desses materiais. Renvoize e colaboradores (1998) [5] propuseram a aplicação de análises estatísticas da morfologia, aliada a outras informações, como forma de proporcionar um sistema razoável de classificação para o gênero *Brachiaria*. Contudo, ainda hoje inexiste uma classificação morfológica precisa em razão da sua variabilidade e estudos moleculares são necessários para auxiliar na classificação das espécies do gênero.

A grande proximidade morfológica das espécies de *Brachiaria* com as espécies do gênero *Panicum,* por exemplo, tem levando diferentes autores a classificar uma mesma espécie em um gênero ou outro, o que dá suporte a uma idéia sustentada por alguns de que *Brachiaria* evoluiu de *Panicum* [10]. Da mesma forma, uma análise filogenética recente concluiu que *Brachiaria* e *Urochloa* formam um grupo monofilético (junto com *Eriochloa* e *Melinis*) e que pesquisas tanto ao nível molecular como morfológicas são necessárias para estabelecer relações claras entre gêneros e espécies deste grupo [3].

Pesquisadores de países como a Austrália e Estados Unidos reclassificaram quase todas as espécies de *Brachiaria* para o gênero *Urochloa,* seguindo trabalhos de Webster (1987) [11], Morone & Zuloaga (1993) [12] e, posteriormente, Gonzalez & Morton (2005) [3]. Porém, as evidências apresentadas nos trabalhos acima ainda conservam controvérsias e não explicam contundentemente as diferenças visíveis, por exemplo, entre *Panicum maximum* e *Brachiaria decumbens*, colocando-os sob o mesmo gênero *Urochloa*. Além disso, os trabalhos mais recentes sugerem a necessidade de estudos mais aprofundados, inclusive

usando marcadores moleculares, para melhor entender as relações entre essas espécies e gêneros. No Brasil ainda se conserva a denominação *Brachiaria* até que novos estudos sejam conduzidos e encontre-se justificativa inquestionável para proceder a mudanças. (http://www.diadecampo.com.br/zpublisher/materias/Materia.asp?id=22378&secao=Colunas%20e%20Artigos)

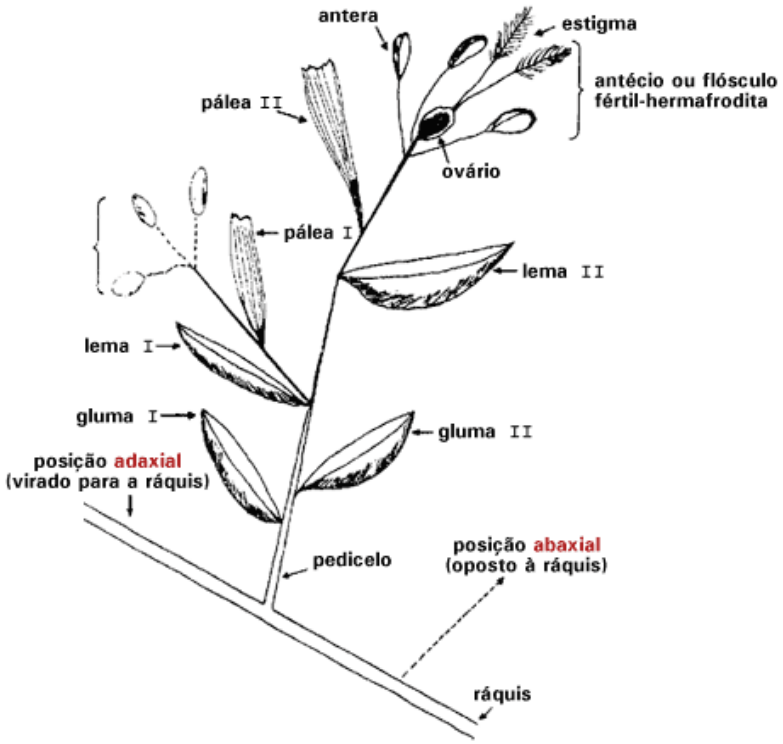


Figura 2- Espigueta com gluma II e antécio hermafrodita abaxial e gluma I e antécio I neutro adaxial, típicos do gênero *Brachiaria* e ráquis variando entre 1,5 a 3 mm de largura Fonte: Rosengurtt (1970) [1].

Diferenciação entre espécies do gênero Brachiaria

Problemas relacionados com classificações incorretas são frequentes entre as espécies de *Brachiaria* comumente utilizadas nas pastagens e entre os acessos de coleções de germoplasma. O intenso intercâmbio de germoplasma também tem causado certa confusão sobre a identidade dos acessos. Diversos estudos [5, 10, 13] destacaram a necessidade de classificar acessos e discriminar espécies corretamente, inclusive para que os bancos de

germoplasma possam ser utilizados com eficiência no melhoramento genético desse gênero.

Como existe grande variabilidade natural entre indivíduos nas espécies de *Brachiaria*, identificar características morfológicas realmente discriminantes torna-se uma difícil tarefa. Renvoize et al. (1998) [5], ao promoverem o agrupamento de 83 espécies de *Brachiaria*, enfatizaram a dificuldade em eleger as características de maior importância na discriminação, sendo a escolha feita, em grande parte, de forma arbitrária e de acordo com a experiência dos próprios pesquisadores. Assis (2003) [9] estabeleceu funções discriminantes para seis espécies de *Brachiaria* baseadas na inclusão simultânea de 24 caracteres morfológicos. Loch (1977) [13] comenta que a comparação de acessos de mesmo nome de dois diferentes locais não garante similaridade, da mesma forma que acessos com nomes diferentes de mesma procedência não garante diferença entre os materiais. *B. decumbens* foi originalmente introduzida no Brasil, em 1952, com o nome de *B. brizantha. B. humidicola* é tratada muitas vezes como sinônimo de *B. dictyoneura* [10]. Renvoize et al. (1998) [5] sustentam que *B. decumbens* cv. Basilisk pertence, na verdade, à espécie *B. brizantha* [9].

A chave proposta por Sendulsky (1977) [2] descreve as dez espécies encontradas com maior freqüência em nosso país e destaca a diferenciação morfológica de *B. ruziziensis* das demais espécies, principalmente através da característica ráquis de 4 mm de largura, e das suas densas nervuras formando um desenho listrado com cloração das folhas verdes amareladas. Já *B. brizantha* possui ráquis de 1 mm de largura, de 2 a 12 racemos longos, de 10 a 20 cm de comprimento, e a primeira gluma com 1/3 do comprimento da espigueta geralmente com uma única série ao longo da ráquis. *B. humidicola* apresenta ráquis de 1mm, espiguetas de até 5 mm de comprimento de contorno arredondado. *B. decumbens* diferencia-se morfologicamente das demais espécies pelas espiguetas com pelos na parte apical, sendo que as duas amostras desta espécie introduzidas no Brasil são diferentes em termos de altura e características das folhas. Por sua vez a chave taxonômica proposta em 1982 pela Royal Botanic Gardens, Kew, Inglaterra (http://www.kew.org/), sugere que *B. ruziziensis* é uma segregação natural de *B. decumbens* e a diferenciação pode ser feita pela largura da ráquis entre 2 a 3,5 mm, enquanto que *B. decumbens* possui a largura de ráquis entre 1,7 e 2 mm, informação confirmada posteriomente por Clayton, W.D. et al. (2006) [14]. J. Gabriel Sánchez-Ken (2012) [15] propôs uma chave taxonômica para o gênero *Urochloa*, na qual também destaca a largura da ráquis de 2,5 a 3mm como diferencial para caracterização de *B. ruziziensis* em relação a *B. humidicola* e *B. brizantha*. Estas últimas diferenciam-se entre si pelo tamanho da gluma, que varia de 3,5 a 5 mm *em B. humidicola*, e até 3,2 mm em *B. brizantha*.

Sistema reprodutivo, ploidia e tamanho do genoma de espécies do gênero Brachiaria

Estudos do sistema reprodutivo de *Brachiaria* identificaram apomixia em diversas espécies [16-18]. A reprodução sexual (anfimixia) é substituída ou combinada com a reprodução assexuada (apomixia) em diversas famílias de angiospermas. A apomixia é entendida como uma forma assexuada de reprodução da planta por meio de sementes. A progênie resultante da reprodução apomítica de uma única planta é clonal, isto é, os indivíduos da progênie são geneticamente idênticos entre si, e também idênticos à planta-mãe. Se a apomixia for obrigatória, torna-se um grande obstáculo para a recombinação genética.

Um padrão frequentemente observado em espécies diplóides de *Brachiaria* é a reprodução sexual. Por outro lado, níveis variados de apomixia são encontrados em espécies poliplóides de *Brachiaria*. As espécies de *Brachiaria* são predominantemente apomíticas facultativas e tetraploides [19]. A conclusão de que são apomíticas facultativas baseia-se na identificação de sacos embrionários típicos de plantas sexuais e apomíticas nesses acessos. Penteado et al. (2000) [20] estimaram níveis de ploidia por citometria de fluxo em uma coleção de germoplasma contendo 435 acessos de braquiária, pertencentes a 13 espécies e observaram vários níveis de ploidia para as diferentes espécies, alguns até então não descritos na literatura científica, como o caso dos pentaplóides. Observou-se também grande variação nas quantidades de DNA total, tanto entre espécies como entre acessos dentro de espécies, sendo *B. brizantha* a espécie mais variável. Ishagaki e colaboradores (2010) [21] estimaram o tamanho do genoma das espécies *B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola* em aproximadamente 615, 1.404, 1.633 e 1.953 Mbp, respectivamente, sendo a primeira espécie diplóide, as duas seguintes tetraplóides e a última hexaplóide. Estes estudos indicaram que o tamanho do genoma está relacionado com a ploidia e também com o modo de reprodução, sexuada ou apomítica. *B. ruziziensis* tem o menor genoma dentre as espécies avaliadas, e é diploide com reprodução sexuada [22, 23]. Por sua vez, *Brachiaria decumbens* e *B. brizantha* são tetraploides (2n = 4x = 36) e apomíticos.

Origem e distribuição das espécies de Brachiaria

Apesar da existência de diversas espécies originárias da América, Ásia e Austrália, a maioria das espécies descritas e cultivadas de braquiária é originária da África, inclusive as

quatro principais espécies cultivadas no Brasil (*B. decumbens, B. brizantha, B. humidicola* e *B. ruziziensis*).

O gênero *Brachiaria* inclui 97 espécies, que podem ser encontradas em climas tropicais e subtropicais, na África e nas Américas [24]. Algumas espécies de *Brachiaria* foram provavelmente introduzidas involuntariamente nas Américas no período colonial, a partir de navios negreiros durante o tráfico de escravos. Sendulsky (1977) [2] relata que as espécies do gênero *Brachiaria* têm sua distribuição nas regiões tropicais de ambos os hemisférios do globo, ocorrendo principalmente na África. No Brasil, até o momento, são conhecidas 15 espécies deste gênero, das quais apenas cinco são consideradas nativas, três foram provavelmente introduzidas há várias décadas, e sete foram introduzidas recentemente, sendo cultivadas como forrageiras. De acordo com Sendulsky (1977) [2], os levantamentos efetuados no Brasil indicaram as 15 espécies relacionadas a seguir:

a) Espécies introduzidas no Brasil

Brachiaria brizantha (Hochst) Stapf
Brachiaria decumbens - sementes da Austrália
Brachiaria decumbens - introdução IPEAN
Brachiaria dictyoneura (Fig & De Mot) Stapf
Brachiaria humidicola (Rendel) Schuwnickerdt
Brachiaria radicans Napper
Brachiaria ruziziensis Germain & Evrard
Brachiaria vittata Stapf

b) Espécies introduzidas no Brasil, provavelmente há dezenas de anos:

Brachiaria extensa Chase
Brachiaria purpurascens (Henr. Blumea)
Brachiaria plantaginea (Link) Hitch

c) Espécies nativas:

Brachiaria adspersa (Trin) Parodi
Brachiaria fasciculata (Se) Parodi
Brachiaria mollis (Sw) Parodi
Brachiaria reptans (L) Gardner & Hubbard
Brachiaria venezuelae (Hack) Heur

Atualmente, a braquiária é a gramínea tropical mais utilizada nas Américas Central e do Sul na produção forrageira. Espécies originárias da Ásia e da Austrália são citadas na literatura, mas poucos estudos abordam estas espécies, suas características agronômicas e biológicas, e importância econômica [25].

Importância econômica da Brachiaria

Cultivares de *Brachiaria* têm impactado a economia de vários países por causa de sua capacidade de crescer em solo infértil com acidez elevada, e ainda ser capaz de produzir forragem altamente nutritiva para ruminantes. Grandes extensões dos trópicos foram convertidos em pastagens a fim de apoiar a pecuária, especialmente na região neotropical. Na América Central, por exemplo, observa-se que o México tem feito grandes esforços para melhorar os cultivares de *Brachiaria,* o que muito incentivou a indústria bovina de carne e leite naquele país. Outros países da América Central também atingiram altos volumes de sementes vendidas e área plantada [26]. No Brasil, até 2004, cerca de 80 milhões de hectares de habitat natural já haviam sido convertidos em pastagem com forrageiras [27] e em 2010 a área plantada com forrageiras no Brasil foi estimada em 101.437.409 hectares [28].

A introdução de braquiária no Brasil provocou uma verdadeira revolução na produtividade das pastagens e na atividade pecuária [29]. Para salientar a importância dos pastos para a economia brasileira, deve ser observado que o Brasil possui o maior rebanho bovino do mundo (180 milhões de cabeças), é o maior exportador de carne bovina e um dos maiores produtores de leite do planeta. Nos últimos anos, o cultivo de *Brachiaria* tornou-se um dos principais componentes das pastagens semeadas com maior área plantada e, portanto, a mais importante neste segmento do agronegócio brasileiro. Estima-se que a área plantada no país com as quatro principais espécies de braquiária (*B. brizantha, B. decumbens, B. ruziziensis* e *B. humidicola*) representa 85% da área coberta com forragens cultivadas [29]. Estima-se que mais de 60 milhões de hectares são cultivados com um único clone de *B. brizantha* (variedade Marandú ou Brizantão) [30]. Isto equivale a uma área significativamente superior à soma da área plantada com as principais culturas agrícolas no país (soja, milho, arroz, algodão, sorgo, feijão, etc) *(*www.conab.gov.br*).* Trata-se, provavelmente, da maior área de monocultura clonal do mundo.

Situação similar é observada com outros ~10 milhões de hectares, plantados com variedades de três outras espécies (*B. decumbens, B. humidicola* e *B. ruziziensis*). Apenas

uma ou duas variedades de cada espécie estão disponíveis para serem usadas no plantio comercial [31]. Isto indica uma situação de risco para a pecuária brasileira, devido à vulnerabilidade genética causada pelo uso em escala de poucos clones de braquiária em grandes extensões territoriais.

Vulnerabilidade Genética da Brachiaria no Brasil

Conforme observado anteriormente, os pastos de braquiária têm papel fundamental na sustentação da pecuária brasileira. Mas deve ser enfatizado que é uma contradição observar que enquanto a área plantada com braquiária no Brasil é continental, a base genética dos pastos plantados é extremamente estreita. Isto coloca os pastos brasileiros, base da alimentação para a produção de carne e leite para consumo interno e exportação, em uma situação ímpar de vulnerabilidade genética.

Levando-se em consideração que apenas uma pequena parcela da produção de carne é destinada à exportação, e que o país apresenta índices zootécnicos considerados baixos em comparação com outros países, ou seja, ainda não atingiu o ponto ideal de equilíbrio entre o resultado técnico e econômico, esta situação de vulnerabilidade tende a se agravar ainda mais nos próximos anos com a expansão da pecuária, se não houver ampla diversificação dos pastos plantados. Isto porque há espaço para o crescimento deste setor no Brasil, ao contrário de outros países exportadores, já que nesses a expansão da pecuária está próxima ao limite de crescimento [29]. Apenas os atuais programas de recuperação de pastagens degradadas no país estimam o replantio de 15 milhões de hectares com forrageiras nos próximos 10 anos, sem levar em consideração a potencial abertura de novas áreas.

O Brasil é também o maior produtor e exportador de sementes de espécies forrageiras tropicais, um mercado que alavanca centenas de milhões de reais por ano em vendas de sementes (ABRASEM, 2005 www.abrasem.com.br). Aliado a técnicas de ILPF (Integração Lavoura, Pecuária e Floresta), a produção de forrageiras tem pela frente um enorme potencial de crescimento em produtividade e qualidade nos próximos anos. Neste cenário, a diversificação dos pastos plantados é de suma importância para o país.

A vulnerabilidade genética detectada nos pastos brasileiros representa um alto risco para o setor agropecuário: estresses bióticos ou abióticos em grandes proporções podem causar prejuízos à produção se não houver diversificação genética das cultivares de forrageiras plantadas no Brasil o mais rapidamente possível. O combate à vulnerabilidade genética deve ser baseado na geração e aproveitamento da diversidade genética oriunda dos

bancos de germoplasma e dos programas de melhoramento para o desenvolvimento de novas cultivares de braquiária.

A espécie Brachiaria ruziziensis

Nome Científico: *Brachiaria ruziziensis* (R. Germ. and C.M. Evrard).
Sinônimo: *Urochloa ruziziensis* (R. Germ. and C.M. Evrard) Crins.

Brachiaria ruziziensis, é também conhecida por "Congo signal grass", "Congo grass", "Ruzi grass", "ruzigrass" e "Kennedy Ruzi grass". As características morfológicas da *Brachiaria ruziziensis* descrevem uma planta perene, rasteira, formando tufos com uma densa cobertura de folhas crescendo a 1-1,5 m de altura, com a base decumbente, tendo espiguetas em 1 ou 2 linhas de um lado da ráquis. As espiguetas são peludas de ~5 mm de comprimento, pilosas na parte apical, bisseriadas ao longo da ráquis. A gluma inferior tem 3 mm de comprimento e surge 0,5 a 1 mm abaixo da espigueta. Ruziziensis apresenta rizomas curtos e fortes, em forma de tubérculos arredondados e com até 15 mm de diâmetro, talo piloso, folhas lineares e lanceoladas, com 100-200 mm de comprimento e 15 mm de largura, de cor verde claro, inflorescência formada por 3-6 racemos de 4-10 mm de comprimento em fita e plana, com floração nos meses de dezembro e janeiro no hemisfério sul [32]. No Brasil, observa-se florescimento nos meses de abril e maio nos estados de Goiás, Minas Gerais e Bahia.

A espécie *B. ruziziensis* está intimamente relacionada com *B. decumbens*, sendo diferenciadas morfologicamente na forma da ráquis, que é subfoliolar e de 2 a 3,5 mm de largura em *B. ruziziensis* e plana variando de 1-1,7 mm em *B. decumbens*. Essa é a principal característica que permite diferenciar morfologicamente ruziziensis das demais espécies de braquiária, apresentada em chaves taxonômicas. Além disto, por apresentar porte maior, possui a gluma inferior 0,5-1 mm distante do resto da espigueta em *B. ruzizensis* em comparação com *B. decumbens* [2]. Em comparação com *B. brizantha* a altura maior e também a largura da ráquis são características de diferenciação morfológicas (Figura 2).

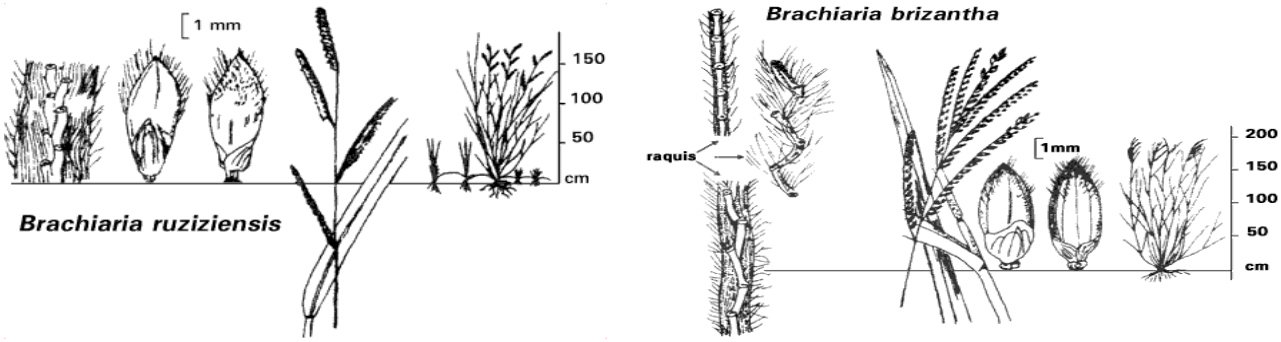


Figura 2 - Detalhe das características da *B. ruziziensis* (esquerda) mostrando a inflorescência formada por 3 a 6 racemos de 4 a10 mm de comprimento. A ráquis largamente alada, com até 4 mm de largura, geralmente de cor arroxeada. As espiguetas de 5 mm de comprimento, pilosas na parte apical, bisseriadas ao longo da ráquis. A altura pode chegar a 1,5 m. *B. brizantha* (direita) com mais rácemos e detalhe da ráquis muito mais fina e planta com altura maior. Fonte: Sendulsky (1977) [2].

A espécie é originária da África, onde ocorre em condições úmidas e não inundáveis, tendo sido encontrada no Zaire, Burundi e oeste do Kenya. Foi cultivada inicialmente no Congo (Zaire), onde junto com *Setaria anceps*, forma a base das pastagens cultivadas. Segundo Serrão & Simão Neto (1971)[33] esta espécie emana um odor peculiar, semelhante ao capim gordura (*Melinis minutiflora* Beauv.).

A *Brachiaria ruziziensis* é uma forrageira de alta qualidade nutricional, com potencial para uso na diversificação das pastagens brasileiras. No Brasil, foi introduzida na década de 1960, vinda da Austrália, embora seja originária da África. O seu plantio pode ser realizado desde o nível do mar até 1.800 m de altitude, nas latitudes de 0 a 25 graus norte ou sul. Essa planta possui muito boa palatabilidade e digestibilidade, é bem precoce, com boa velocidade de rebrota, níveis de proteína variáveis entre 11 e 13%, de acordo com as estações do ano. É indicada especialmente para bovinos, embora equinos, ovinos e caprinos a consumam. O crescimento é rápido no início da temporada de chuvas e apresenta compatibilidade no consórcio com leguminosas. Muitos agricultores têm utilizado a *B. ruziziensis* em áreas de cultivo de soja para cobertura vegetal, no período de entressafra da cultura e como pasto no inverno. A espécie é bem adaptável para sistemas de integração lavoura-pecuária-floresta como cobertura de solo para o plantio direto, com menos exigências de herbicida para dessecação (Figura 3).

Essa planta se comporta bem em solos de fertilidade média a alta, tem razoável

tolerância ao frio, baixa tolerância à umidade e média tolerância à seca. Apresenta excelente velocidade de recuperação após as primeiras chuvas, no final da seca, o que lhe confere bom destaque para plantio nas diversas regiões do Brasil.

A praga mais problemática para o cultivo de pastos de *B. ruziziensis* é a cigarrinha, que é uma praga conhecida dos trópicos [24]. As cigarrinhas são insetos sugadores que, durante o período da seca, permanecem na pastagem na fase de ovo, pois não encontram condições favoráveis para a eclosão. Com o início do período chuvoso estes ovos eclodem e dão origem às ninfas (formas jovens das cigarrinhas)

Além do calor, as cigarrinhas dependem, para o seu desenvolvimento, de muita umidade. Isto é facilmente notado, uma vez que as ninfas, geralmente localizadas na base das plantas, vivem no interior de massas de espuma por elas secretadas. Tem havido um grande esforço no sentido de se identificar gramíneas resistentes às cigarrinhas. Entre as *Brachiarias, B. decumbens* cv. Basilisk e *B. ruziziensis* foram consideradas susceptíveis, enquanto *B. humidicola*, tolerante (sofre menos danos do que outras *Brachiarias*) [34].. Altas infestações desses insetos têm influência direta na qualidade das gramíneas e promovem a redução drástica da capacidade de suporte das mesmas.

Figura 3 – *Brachiaria ruziziensis* utilizada em consórcio com milho em sistema de integração lavoura-pecuária.

As braquiárias também são atacadas por formigas cortadeiras e por um conjunto de doenças fúngicas, bacterianas e virais. Daí a necessidade de desenvolvimento de cultivares resistentes a doenças como uma alternativa de baixo custo para o controle químico e controle preventivo de doenças e pragas nos pastos [35].

Dentre as quatro espécies de *Brachiaria* mais cultivadas no Brasil, a *Brachiaria ruziziensis* (2n = 2x = 18) destaca-se por ser uma espécie diplóide com reprodução sexuada, uma vantagem para o uso de métodos de melhoramento com vistas ao desenvolvimento de novas cultivares. Este ainda é um desafio em espécies tetraplóides, que normalmente apresentam reprodução apomítica, o que impede o desenvolvimento de novas cultivares através de recombinação gênica.

Note-se que após a tetraploidização, a *B. ruziziensis* pode ser cruzada com outras espécies de *Brachiaria* de interesse agrícola como *B. decumbens* e *B. brizantha*, tornando a

introgressão inter-específica de genes possível. O melhoramento genético de *B. ruziziensis* pode ser usado, portanto, como base para alavancar a diversificação e melhoramento das espécies poliploides através da duplicação cromossômica de genótipos superiores obtidos por recombinação, e posterior cruzamento com acessos de *B. decumbens* ou *B. brizantha*. Mas é importante destacar que o melhoramento genético de *B. ruziziensis* por si só apresenta grande potencial para a pecuária brasileira, contribuindo para o desenvolvimento de variedades mais produtivas e nutritivas, com produção de sementes de alta qualidade, incluindo os híbridos, transformando *B. ruziziensis* em uma cultura essencialmente agrícola, especialmente para a rotação de culturas em sistemas integração lavoura-pecuária e floresta (ILPF).

A importância da B. ruziziensis para os programas de melhoramento

Brachiaria ruziziensis pode contribuir para a diversificação genética dos pastos brasileiros, promovida por recombinação gênica em nível diplóide. Ao mesmo tempo, genótipos superiores do programa de melhoramento de *B. ruziziensis* podem ser potencialmente utilizados, após tetraploidização, em programas de melhoramento de espécies poliplóides.

Entre as espécies de braquiária, *B. ruziziensis* é a que apresenta maior qualidade forrageira [36] e grande aptidão para emprego em sistemas ILPF. Pelo valor que apresenta em sistemas ILPF, *B. ruziziensis* pode ser considerada não como uma forrageira tradicional, mas potencialmente como uma nova espécie agrícola, para uso em rotação de culturas neste sistema. Para isso, no entanto, é necessário uma maior tecnificação na produção de sementes, no plantio, no manejo e no melhoramento genético, visando aumentar a produtividade e a qualidade de forragem para consumo verde ou como feno. Destaque-se que, por apresentar sementes deiscentes, nenhuma espécie de braquiária pode ser considerada domesticada. A domesticação de braquiária é um passo fundamental a ser dado pelos programas de melhoramento genético.

Juntas, *B. ruziziensis*, *B. decumbens* e *B. brizantha* são as forrageiras mais importantes na América do Sul [37]. Em *B. brizantha*, a apomixia é o modo predominante de reprodução, com apenas um acesso com reprodução sexual descrito entre 275 analisados na coleção de germoplasma do Brasil [20]. A existência de diplóides sexuais em *B. ruziziensis* e, potencialmente em outras espécies de braquiária, abre a possibilidade de

melhoria do *pool* genético sexual do gênero *Brachiaria*, vertente ainda pouco explorada. Pode-se dizer que a melhoria das populações com reprodução sexual deve ser considerada essencial para qualquer programa de melhoramento de espécies do gênero. Neste contexto, a possibilidade de cruzamento com *B. ruziziensis* apresenta-se também como opção relevante para a diversificação dos genótipos de espécies como *B. decumbens* e *B. brizantha*.

A escolha da *B. ruziziensis* como espécie de referência para o melhoramento pode ser sintetizada por dez importantes motivos abaixo relacionados:

1. Reprodução sexual: permite recombinação, seleção e desenvolvimento de linhagens superiores;
2. Diversidade genética: base do melhoramento, B. ruziziensis possui germoplasma com expressiva diversidade genética;
3. Métodos convencionais de melhoramento podem ser usados (vantagem adicional: possibilidade de clonagem);
4. Genoma pequeno: 600 Mpb – genética molecular e genômica é facilitada pelo menor tamanho do genoma;
5. Aptidão: ILPF e pasto;
6. Área plantada em franca expansão (especialmente ILPF);
7. Boa qualidade nutricional (mesmo sem melhoramento);
8. Maior uniformidade no florescimento e produção de sementes;
9. Proximidade genética com *B. brizantha* e *B. decumbens*;
10. Ponte para melhoramento genético de outras espécies através de poliploidização.

Não há informação genômica disponível para o gênero Brachiaria

Embora tenha grande importância na atividade agropecuária e na economia de vários países, há uma falta generalizada de informação sobre os genomas das espécies de *Brachiaria*. Pouco ou nada se sabe sobre o número e composição gênica, distribuição de famílias de genes, abundância e diversidade de retro-elementos, localização de QTLs para características economicamente importantes, colinearidade dos genomas das várias espécies de braquiária, ortologia com espécies modelo e abundância de sequências repetitivas. A pouca disponibilidade de ferramentas genômicas, como marcadores moleculares microssatélites e SNPs para serem utilizados em apoio aos programas de melhoramento é evidente [38].

O programa de melhoramento *B. ruziziensis* pode ser intensamente reforçado se houver o emprego de ferramentas de genômica para apoiar a seleção de genótipos superiores. Isso certamente vai favorecer um desenvolvimento mais dinâmico de novas cultivares para esta espécie. Como consequência, a disponibilidade de informação genômica de *B. ruziziensis* terá forte impacto na eficiência dos programas de melhoramento genético.

Comparado a outros genomas de gramíneas, *B. ruziziensis* tem um genoma relativamente pequeno (~ 600 Mpb) [21] , semelhante a outros modelos de espécies de cereais, tais como arroz (430 Mpb) e sorgo (700 Mpb). De certa forma, esta característica viabiliza as iniciativas de análises genômicas de sequenciamento e montagem, tendo em conta a utilização de recursos de sequenciamento de última geração, com vistas ao desenvolvimento de ferramentas moleculares para apoiar programas de melhoramento. Espécies tetraploides de *Brachiaria* (por exemplo, *B. decumbens*, *B. brizantha*) têm genomas maiores e mais complexos (> 1.600 Mpb), o que dificulta e encarece significativamente a processo de sequenciamento, montagem *de novo* e análise do genoma. O sequenciamento em larga escala do genoma possibilita aumentar significativamente o conhecimento do genoma desta espécie.

O avanço do conhecimento de genomas modelo como o de arroz e o advento de técnicas de sequenciamento de nova geração (*Next Generation Sequencing*) abrem a possibilidade de grande salto no conhecimento do genoma de espécies relativamente pouco conhecidas. Os recentes desenvolvimentos de tecnologias de sequenciamento de alto desempenho (UHT - *Ultra High Throughput*) a custos acessíveis permitem hoje propor experimentos que até poucos anos atrás eram impensáveis [39], como o sequenciamento de genomas inteiros de uma ou mais variedades de uma espécie, seguido do alinhamento destes genomas, identificação de regiões polimórficas e, finalmente, associação do polimorfismo de DNA à variação fenotípica.

Estas novas tecnologias de sequenciamento em larga escala, usadas em conjunto com ferramentas computacionais de bioinformática, constituem um poderoso recurso para a compreensão sistemática dos genomas, especialmente de espécies ainda pouco estudadas e de alto impacto econômico. Portanto, cabe a proposta de desenvolvimento de metodologias e uso de recursos de bioinformática em experimentos de análise e de sequenciamento genômico em larga escala, para o emprego de marcadores SSRs, SNPs e indels (inserções/deleções), com foco no desenvolvimento de ferramentas genômicas para seleção assistida por marcadores moleculares no programa de melhoramento genético da *B. ruziziensis*. No momento, não há nenhuma ferramenta genômica (ex. marcadores moleculares ou painéis de

genotipagem em escala) disponível para emprego no melhoramento genético de *B. ruziziensis*.

O Sequenciamento de DNA em larga escala

O sequenciamento tradicional (método Sanger)

Em 1977, o premiado cientista inglês Fred Sanger (prêmio Nobel por duas vezes) descreveu uma metodologia para determinação das sequências de DNA, posteriormente denominado método Sanger ou dideoxi [40]. O método Sanger foi virtualmente o único método de sequenciamento de ácidos nucléicos utilizado nas três décadas seguintes [41], servindo de base para a era genômica na Biologia. Este período caracterizou-se por avanços técnicos, principalmente pela automatização de equipamentos de sequenciamento e análise de segmentos de DNA, que permitiram o sequenciamento de um grande número de genomas completos de diversos organismos.

Porém, apesar do enorme sucesso obtido, a necessidade de sequenciamento com menores custos, maior rapidez e maior eficiência ainda estariam por vir. Novas estratégias de sequenciamento de DNA foram desenvolvidas [39] ao final deste período e os resultados obtidos por estas novas tecnologias projetaram uma revolução na Biologia, pois o volume de dados gerado é de duas a três ordens de magnitude maior que os obtidos pela tecnologia Sanger, e a um custo bem inferior.

Nos primeiros projetos de sequenciamento genômico com a tecnologia Sanger, os fragmentos sequenciados eram caracterizados por um número pequeno de sequências com tamanho de até 1000 de bases, cujo processamento era realizado em períodos de semanas e meses em extenuante trabalho em laboratórios. No início do novo milênio, porém, a automatização das tecnologias de sequenciamento foi aperfeiçoada até um ponto em que equipamentos totalmente automatizados passaram a gerar sequências de um modo paralelizado durante 24 horas por dia. Grandes centros de sequenciamento de genomas ao redor do mundo abrigaram dezenas dessas máquinas de sequenciamento. Isto, por sua vez, levou à necessidade de criação de novos algoritmos montadores de genoma, utilizando sequências variando de 35 a 1000 bases de comprimento, e com taxas de erro de sequenciamento variando de 0,5 e 15%, que ainda podem conter artefatos complexos como repetições.

Com a tecnologia Sanger, projetos de sequenciamento de bactérias com 20.000 a 200.000 segmentos de leitura podiam ser montados em um computador. Os maiores, como o genoma humano, com cerca de 3 bilhões de bases, já necessitavam de grandes estruturas de computação para processamento dos dados, envolvendo vários laboratórios e centros de pesquisa.

As novas tecnologias NGS

Em meados de 2005, uma nova técnica para sequenciar segmentos de DNA foi apresentada pela companhia 454 Life Sciences[1] (posteriormente Roche), que consistia na paralelização do processamento de sequenciamento, utilizando a nanotecnologia e a metodologia de pirosequenciamento. As principais vantagens apontadas para o uso desta tecnologia eram a rapidez, o volume de sequências geradas e a facilidade técnica por contornar a necessidade de clonagem de fragmentos de DNA.

No sequenciamento 454 era possível obter de 400 a 500 mil sequências em cada corrida (ciclos de sequenciamento em paralelo), culminando em aproximadamente 100 Mb sequenciados em poucas horas de trabalho [42]. O pirosequenciamento trouxe a viabilidade comercial para a tecnologia 454 em relação ao método Sanger. Este novo método de sequenciamento em escala gerava, por outro lado, segmentos de leitura muito mais curtos, inicialmente cerca de 100 bases. Atualmente, esta tecnologia permite obter 1.000.000 de segmentos de leitura por corrida com tamanho aproximado de 1.000 pb (GS FLX). Esta metodologia tem como principais desvantagens o alto custo dos reagentes e a taxa relativamente elevada de erro [43]. No entanto, devido à capacidade muito maior de geração dados de sequenciamento e custo menor do que sequenciamento Sanger, a adoção desta tecnologia por centros de genoma gerou o desafio de desenvolvimento de programas de bioinformática montadores de sequências para tratar desse novo tipo de informação. Logo em seguida, surgiram novas tecnologias, batizadas de sistemas de sequenciamento paralelo em massa de ultra-desempenho, ou *ultra-high throughput sequencing* [44]. Desde 2006, a Illumina Inc. tornou disponível esta nova tecnologia, capaz de gerar cerca de 100 milhões de segmentos de leitura por corrida. O procedimento estava inicialmente limitado a produzir sequencias com um comprimento de apenas 36 bases, tornando-o menos adequado para a montagem *de novo* de genomas. Contudo, foram surgindo novos equipamentos como o Genome Analyzer (Illumina GA) da geração de 2011, uma das tecnologias mais utilizadas

[1] http://www.454.com

recentemente, que emprega tecnologia SBS - *Sequencing By Synthesis*, capaz de gerar até 600 Gb com segmentos de 76 pares de base em média. Equipamentos ainda mais recentes da tecnologia permitem leitura de segmentos de DNA 100 bases em média (*www.illumina.com*).

Outras tecnologias, como a SOLiD da Applied Biosystems, foram disponibilizadas no período, e tecnologias mais recentes, como IonTorrent e PacBio, continuam a avançar na capacidade de sequenciamento em escala. Outros exemplos incluem: (a) equipamentos sequenciadores de DNA, como o SOLiD[2] (Applied Biosystems, atual Life Technologies), que apresentam uma capacidade de geração de 80 a 160 gigabases de sequências de DNA por corrida com o tamanho de segmentos de leitura de 50 pb; (b) Helicos Biosciences, que propõe uma metodologia de sequenciamento em escala que utiliza os fragmentos de DNA com poli-A adicionado a adaptadores na cauda que estão ligados à superfície da célula de fluxo. O protocolo envolve extensão e sequenciamento com lavagens cíclicas da célula de fluxo com nucleotídeos marcados com fluorescência. As leituras são curtas, até 55 bases por corrida [45]; (c) Pacific Biosciences, que propõe tecnologia que permite a leitura de segmentos de até 15.000 nucleotídeos, com média de comprimentos de leitura de 2,5-2,9Kb (http://www.pacificbiosciences.com).

Significativo avanço foi alcaçado no sentido de aumentar a quantidade e qualidade das sequências de DNA, bem como a capacidade de montagem de genomas completos. As atuais tecnologias disponíveis, é claro, apresentam vantagens e desvantagens. Na avaliação da performance de cada tecnologia, geralmente considera-se todas as etapas do sequenciamento, inclusive cuidados com a preparação de amostras, sequenciamento *per se*, tratamento de imagem, e análise de dados. Por exemplo, a maioria das abordagens de sequenciamento de DNA possui um passo de clonagem *in vitro* para amplificar moléculas de DNA individuais, porque os métodos de detecção molecular não são suficientemente eficientes para a detecção de molécula única. A PCR em emulsão, uma das técnicas mais utilizadas, possibilita o isolamento de moléculas individuais de DNA, juntamente com esferas (*beads)* revestidas em gotículas aquosas dentro de uma fase de óleo, seguida da reação em cadeia da polimerase (PCR). Cada uma das esferas fica revestida com cópias clonais da molécula de DNA, seguida de imobilização dos *beads* para mais tarde serem submetidas ao sequenciamento. PCR em emulsão é usada nos métodos desenvolvidos por Margulis et al. (2005) [46] (Roche 454), Shendure et al. (2005) [47] (também conhecido como sequenciamento polony) e sequenciamento de sólidos (desenvolvido por Agencourt, depois Applied Biosystems, agora

[2] http://solid.appliedbiosystems.com

Life Technologies). A combinação única de protocolos específicos distingue uma tecnologia da outra e determina o tipo de dados produzidos a partir de cada plataforma. O tratamento destes dados representa um desafio quando comparamos as plataformas quanto à qualidade e custo. Não há consenso na literatura sobre a estimativa de qualidade de dados de sequenciamento em uma plataforma e sua equivalência em outra plataforma [43].

As tecnologias exemplificadas acima são essencialmente complementares [48]. A tecnologia 454 vem sendo utilizada para o sequenciamento *de novo* de genomas procariotos [49], sequenciamento de ESTs [50] e metagenômica [51]. As tecnologias de leituras curtas (Illumina e SOLiD) têm sido utilizadas para o re-seqüenciamento de genomas com base em genoma referência, medição global dos níveis de mRNAs, descoberta de micro RNAs, estrutura de cromatina e análise epigenética [52]. Mais recentemente, a 454 da Roche GS FLX e Illumina / Solexa Genome Analyzer IIx têm sido usados principalmente na montagem *de novo* de transcriptoma. Embora a tecnologia de sequenciamento Roche possa produzir segmentos de leitura mais longos, a plataforma Illumina possibilita a obtenção de cobertura mais profunda e maior precisão com o mesmo custo, o que é benéfico para a descoberta de genes e marcadores moleculares [53].

Outras estratégias atuais de sequenciamento de DNA incluem a rotulação da DNA polimerase [54] e a leitura da sequência de cadeias de DNA através de nanoporos. Incluem ainda técnicas especiais de microscopia, como a microscopia de força atômica ou microscopia eletrônica de transmissão, que são usadas para identificar as posições dos nucleotídeos individuais dentro de fragmentos de DNA longos (> 5000 pb) por marcação de nucleotídeos com os elementos mais pesados (por exemplo, átomos de halogênio) para a detecção visual e de gravação [55]. A decisão para usar uma estratégia ou outra baseia-se na aplicação biológica a qual se destina, bem como custo, esforço, tamanho estimado do genoma, sua complexidade e considerações de tempo [56]. Por exemplo, identificação e catalogação da variação genética em várias cepas de genomas relacionadas, tais como aquelas encontradas em espécies de bactéria, *C. elegans*, e plantas como *Arabidopsis thaliana*, podem ser realizada por NGS, alinhando os segmentos com seus genomas de referência. Esta abordagem é atualmente substancialmente mais barata e mais rápida do que sequenciamento Sanger.

Genômica computacional: o desenvolvimento de ferramentas computacionais é fundamental para o estudo e análise de genomas

A enorme quantidade de dados de sequência de DNA gerados por tecnologia NGS, juntamente com os artefatos e erros inerentes a cada tecnologia de sequenciamento, desafiam os projetos de montagem de genomas completos de diferentes espécies. Há menos de uma década (2004), apenas o montador Newbler era disponível para este fim, aplicado à montagem de fragmentos gerados por sequências produzidas pelo sequenciador Roche 454 (software proprietário). Apresentado em meados de 2007, a versão híbrida do montador MIRA [57] foi o primeiro montador de uso livre desenhado para montar segmentos de 454 e misturas de segmentos 454 e Sanger, utilizando sequências longas de diferentes origens. No final de 2007, o montador SHARCGS [58] foi publicado para montagem de segmentos curtos oriundos da tecnologia Illumina, rapidamente seguido por uma série de outros softwares.

Conforme mencionado anteriormente, as novas tecnologias de sequenciamento de DNA envolvem a paralelização no número de amostras analisadas através de miniaturização de reações, substituindo o sequenciamento em capilares do método Sanger, e incluindo novas químicas no processo. Entretanto, estas tecnologias, que geram milhões de sequências de leituras distintas, têm como característica determinante a produção de fragmentos de sequências menores do que os obtidos com sequenciamento Sanger, em geral entre 35 e 250 bases. Atualmente, para a montagem de genomas existem diferentes opções de montadores, adequados à montagem utilizando diferentes tamanhos de fragmentos, inúmeros formatos de arquivos, e aplicados a genomas de diferentes complexidades. Novos avanços no processo de montagem são esperados com a integração de bancos de dados e com a exploração de múltiplas estratégias de sequenciamento, sempre com o propósito de enfrentar o desafio de montagem de genomas complexos. Os genomas grandes e marcados por abundância de sequências repetitivas ainda constituem um grande desafio para o desenvolvimento de algoritmos de montagem a partir de sequências curtas de DNA. Normalmente, vários montadores são combinados para contornar este problema. Enquanto isso, a precisão e o comprimento dos fragmentos sequenciados vêm aumentando paulatinamente [59].

Da mesma forma, o aumento explosivo na quantidade de informação de sequências através das modernas técnicas de sequenciamento em larga escala requer o desenvolvimento de ferramentas computacionais e algoritmos mais eficientes para a análise dessa imensa quantidade de dados. Para maximizar o potencial de se construir a sequência completa de todos os cromossomos de um organismo, a bioinformática tem um papel fundamental, pois os pequenos fragmentos devem ser remontados para obter a sequência inteira de DNA. Além disso, a bioinformática pode ajudar a transformar informação genética em conhecimento biológico aplicável.

Desde 2004, o Instituto Nacional do Genoma Humano já distribuiu mais de US $ 100 milhões para o desenvolvimento de tecnologias NGS, o que tem promovido o progresso nesta área por meio de vários empreendimentos comerciais. Conforme mencionado anteriormente, várias empresas possuem tecnologias NGS em vários estágios de desenvolvimento e comercialização. Deve ser enfatizado que a produção de bilhões de segmentos de DNA oriundos de tecnologia NGS também requer a infraestrutura de tecnologia da informação para aumentar a eficiência na transferência de dados, controle, armazenamento e análise computacional para alinhamento ou montagem de genomas. Além disso, são requeridos sistemas de gestão de informação para rastreamento de amostra e gestão de processos laboratoriais. Este é o tema do qual trata a chamada genômica computacional [60]. Genômica computacional é o estudo da composição, estrutura e função do material genético dos organismos por meio de recursos computacionais. Avanços em bioinformática com foco na genômica computacional estão em andamento, e as melhorias nestes sistemas são necessárias para manter o ritmo de evolução das tecnologias da NGS. É possível que os custos associados com a manipulação e análise de dados venham, em breve, superar os custos de geração de informação por sequenciamento.

A exploração sistemática de bases de dados gerados nos projetos genoma é um desafio importante para transformar informação em tecnologia. Tecnologias genômicas de alto desempenho que permitem a análise de milhares de genes em paralelo, integradas aos programas de melhoramento, estão abrindo novas perspectivas para a compreensão das relações complexas entre variabilidade genética e diversidade fenotípica e, por fim, a aplicação deste conhecimento na seleção direcional para obtenção de plantas elite. O desafio agora é a exploração sistemática e inteligente deste banco de informações genômicas e os recursos experimentais gerados em paralelo.

Deve ser pontuado, conforme será visto no presente trabalho, que simulações com o genoma de espécies já conhecidas, como o arroz (*Oryza sativa*), permitem, a partir de dados de sequenciamento em escala, redefinir estratégias de montagem *de novo* do genoma de espécies órfãs de informação genômica, como a *B. ruziziensis.*

Montagem "de novo" de genomas x Montagem com genoma de referência

A montagem do genoma consiste em um conjunto de procedimentos em que se busca

organizar um grande número de sequências curtas de DNA em um espaço linear, com o objetivo de representar a molécula de DNA que compõe cada cromossomo da espécie estudada. Em projetos de sequenciamento, todo o DNA de uma fonte (geralmente um único organismo, desde uma bactéria a um mamífero) é primeiro fragmentado em milhões de pedaços pequenos. Estas peças são depois "lidas" por máquinas de sequenciamento automatizadas, que podem decifrar segmentos de leitura ("reads") que, em geral, variam de 76 a mais de 1 Kb de comprimento. Os algoritmos de montagem do genoma funcionam tomando todas estas peças de uma vez, alinhando-as umas às outras, tentando identificar as regiões onde dois segmentos de leitura se sobrepõem. Estas sobreposições podem ser incorporadas linearmente em um processo de montagem, que é contínuo. Quanto mais curtas as sequências, maior a quantidade de sobreposições necessárias para que possa executar esta tarefa. A cobertura genômica, isto é, o número de vezes que uma determinada região do genoma é coberta por segmentos de leitura, contribui para aumentar a acurácia de identificação da sequência de DNA na região considerada. O emprego de segmentos de leitura com pareamento de extremidades, isto é, com identificação de sequência de DNA nas suas duas extremidades ("paired-end reads"), separadas por uma distância de referência, facilita o processo de obtenção de sequências montadas ("contigs").

O emprego de tecnologia NGS no sequenciamento genômicogera facilmente dois ou três bilhões de segmentos de leitura de DNA com 100 cópias cada [61], que podem ser usados na montagem do genoma da espécie. A montagem representa, naturalmente, um desafio de alta complexidade, visto. que as sequências de sequenciamento NGS são pequenas (ex. segmentos de 76 pb gerados por sequenciador Illumina GAIIx)..Durante a montagem de genomas, os fragmentos de leitura geralmente são alinhados com uma sequência genômica reconhecida como "referência para a montagem do genoma". Na ausência de um genoma de referência, as sequências de leitura devem ser usadas para uma montagem *de novo* do genoma. A decisão para usar a estratégia de montagem *de novo* ou baseada na referência, caso esta última esteja disponível, baseia-se na aplicação biológica, no custo, no esforço necessário para atingir a acurácia necessária e considerações de tempo de montagem.

O termo sequenciamento "*de novo*" vem do latim e significa "*desde o princípio".* Refere-se, pois, a métodos utilizados para determinar a sequência de DNA quando não há nenhuma sequência genômica conhecida anteriormente e disponível para uso como referência. As diferentes estratégias de sequenciamento *de novo* têm vantagens e desvantagens em velocidade e precisão quando comparadas entre si. A montagem *de novo* é

quase sempre complexa e difícil, particularmente quando o genoma é grande e o DNA analisado possui sequências que se repetem muitas vezes, causando falhas na montagem.

Principais parâmetros considerados na montagem "de novo" de genomas

A definição de parâmetros e medição da acurácia da montagem *de novo* de um genoma não é tarefa trivial. A tendência é, muitas vezes, otimizar o valor de N50. Este valor é o parâmetro usado para estimar o comprimento dos contigs montados, isto é, o menor comprimento de contig a partir do qual o somatório de todos os contigs representa a metade do comprimento de todos os contigs montados. Otimizar o valor N50 pode fazer com que os contigs se tornem cada vez maiores, mesmo quando há pouca informação se esses contigs são precisos ou não. Neste caso, o alinhamento com BLAST tem sido utilizado em simulações para comparar os contigs montados às sequências de referência, verificar como eles se encaixam no modelo e a quantidade de contigs com deficiência de montagem.

Salzberg e colaboradores (2012) [62] descrevem o desempenho relativo dos diferentes montadores, e observam diferenças significativas na dificuldade de montagem, as quais parecem ser inerentes aos próprios genomas. Os autores concluem que: (a) a qualidade dos dados, e não o montador, tem um efeito dramático sobre a qualidade de um genoma montado; (b) o grau de contiguidade de um conjunto de dados varia muito entre diferentes montadores e genomas diferentes; (c) a correção de uma montagem também varia muito e não está bem correlacionada com estatísticas sobre contiguidade.

Em geral, os montadores *de novo* trabalham com dois algoritmos principais. Os montadores baseados em cadeia de dados e *sobreposição-layout-consenso* (OLC) são bem adaptados para sequências muito curtas de genomas pequenos. Para grandes conjuntos de dados de mais de cem milhões de leituras curtas, *de Bruijn graph* (grafo de Bruijn) parece ser mais apropriado [59]. O *de Bruijn graph* [63] é um algoritmo que quebra os segmentos de leitura em k-mers antes de montá-los em contigs. A abordagem de grafos forma contigs ligando dois fragmentos (k-mers) com k ou mais nucleotídeos sobrepostos. Contudo, ambas as abordagens enfrentam o problema de falso-positivos e de leituras errôneas. Além disso, a falta de vértices do grafo, devido à não uniformidade de cobertura e de segmentos de repetição, pode ser também um fator limitante. A escolha apropriada de k é crucial, mas para qualquer k, há sempre um problema: um k pequeno favorece a situação de leituras errôneas e uma cobertura não uniforme, e um grande k favorece regiões de repetição curtas. A proposta de uma abordagem iterativa (*de Bruijn graph*) de captura de pequenos a grandes k e de todos

os valores entre eles parece ser uma alternativa viável [64].

Em relação aos parâmetros para mapeamento de segmentos de leitura na sequência de referência, tanto a variação dos parâmetros de fração de alinhamento mínimo das sequências de leitura quanto o percentual de identidade tem efeito na eficiência de mapeamento. O efeito de variação da fração de alinhamento é maior na extensão de montagem, porém menor do que o efeito da variação do percentual de identidade na identificação de polimorfismos. Na estringência máxima dos parâmetros, por exemplo, a detecção de variações alélicas não é possível, uma vez que todas as sequências mapeadas devem ter 100% de identidade, e o percentual do genoma montado que é mapeado diminui. A escolha destes parâmetros é importante, pois influencia na cobertura e na tolerância de erro, o que é decisivo para identificação de polimorfismos.

Desafios da montagem "de novo"

Em relação à complexidade do genoma e ao tempo de montagem, montagens *de novo* são ordens de magnitude mais lentas, consomem muito mais memória de processamento e exigem mais interatividade e atenção do que montagens com referência. Isto é principalmente devido ao fato de que o algoritmo de montagem precisa comparar cada leitura a cada segmento diferente em uma operação que tem alta complexidade.

Esta complexidade da montagem de sequências é ocasionada por dois fatores principais: o número de fragmentos e os seus comprimentos. Quanto mais fragmentos melhor a identificação de sobreposições de sequências. Embora as sequências mais curtas sejam mais rápidas para alinhar, elas também complicam a fase de distribuição linear do segmento montado, como na construção de andaimes (*scaffolds*) no realinhamento da orientação e junção dos contigs, buscando segmentos maiores e ligando-os para criar andaimes. Os *scaffolds* são segmentos resultantes do alinhamento final do processo de construção de elementos que possam ligar dois ou mais contigs. O procedimento de *scaffolding* é importante, pois pode aumentar bastante a média do tamanho dos contigs e consequentemente, o N50.

Montagem de genomas é um problema computacional ainda mais complicado quando o genoma considerado contém um grande número de sequências repetitivas idênticas. Estas repetições podem estar distanciadas por milhares de nucleotídeos, e algumas ocorrem em milhares de diferentes locais, especialmente nos grandes genomas de plantas e animais [65], tornando a tarefa de montagem *de novo* especialmente complexa.

Montagens *de novo* têm sido relatadas para genomas bacterianos e de mamíferos [66], mas existem desafios consideráveis para a sua aplicação em grandes genomas. O uso de segmentos de leitura *paired-end* pode, em certa medida, compensar o comprimento de fragmentos de leitura simples (s*ingle-end*). Programas montadores diversos, tais como SSAKE [66], SOAPdenovo [67] e Velvet [68], exploram as informações de sequenciamento *paired-end* com o propósito de aumentar a acurácia dos contigs montados.

Uma estratégia para a melhoria da qualidade do alinhamento ou montagem tem sido aumentar a cobertura genômica. Embora isto pareça razoável, experimentos conduzidos em nosso laboratório com sequências de leitura *paired-end* usando sequenciador Illumina têm mostrado que existe um limite de saturação através do qual se torna muito difícil avançar na extensão linear do genoma apenas com o aumento da cobertura. O Capítulo 3 descreve esta situação na saturação de cobertura em sequencias do genoma cloroplástico.

Uma vez que cada plataforma NGS produz um padrão diferente de sequências de tamanho e cobertura variável, a mistura de tipos diferentes de NGS na montagem pode contribuir para corrigir deficiências. Aury e colaboradores (2008) [69] relatam uma mistura de sequenciamento utilizando as plataformas Roche 454 e Illumina que resultou em melhora nas montagens *de novo* de genomas microbianos em comparação com as montagens de qualquer uma destas plataformas em separadamente.

A otimização dos parâmetros do programa de montagem e a eliminação de sequências de leitura de baixa qualidade também concorrem para a melhoria da montagem de genomas de forma significativa. A validação da montagem pelo uso de diferentes programas com algoritmos alternativos também é uma opção para obter melhores resultados [70].

Originalmente, a maioria dos grandes centros de sequenciamento de DNA desenvolveu seu próprio software para montar as sequências que eles produziram. No entanto, isto foi alterado, pois com o aumento do número de técnicas e de centros de sequenciamento a tarefa dos softwares de montagem tornou-se mais complexa.

Montagem "de novo" e a caracterização de genomas de espécies sem informação genômica

As tecnologias de sequenciamento de nova geração proporcionam uma economia de custo, trabalho e de análise e caracterização de genomas. Embora muitas ferramentas de bioinformática para montagem de genomas tenham sido desenvolvidas para o emprego de sequências curtas de dados (neste caso em torno de 76 pb) para a análise genômica, a aplicação destes recursos para o conhecimento sobre genomas ainda é muito limitada [70] e,

na sua grande maioria, como espécies de *Brachiaria*, ainda inexistente.

O uso de dados de segmentos curtos de leitura para caracterizar o genoma de um organismo com parcos conhecimentos genômicos, juntamente com uma estratégia de montagem *de novo,* representam oportunidade e desafio importantes para o avanço do conhecimento destas espécies [71].

Diversas iniciativas e estratégias de sequenciamento *de novo* de genomas de forma parcial ou completa surgiram recentemente. Um bom exemplo foi a utilização de diferentes tamanhos de fragmentos na construção das bibliotecas, com dimensões de inserção de cerca de 150 pares de bases (bp), 500 pb, 2 kb, 5 kb e 10 k, combinadas com a tecnologia de sequenciamento Illumina para o sequenciamento do genoma do urso Panda [72]. Prevê-se que este tipo de abordagem possa representar uma contribuição significativa para o desenvolvimento de recursos genômicos, para estudos funcionais e para apoiar programas de melhoramento de plantas e animais.

Sequências gênicas (conteúdo gênico) do genoma

A predição computacional de genes, ou descoberta de genes a partir da análise da sequência montada do genoma, refere-se ao processo de identificação de regiões de DNA genômico que codificam sequencias protéicas ou que regulam a atividade gênica. Isto inclui sequências codificadores de proteínas, assim como os genes de RNA, mas pode também incluir a previsão de outros elementos funcionais, tais como regiões reguladoras. Predição de genes é um dos primeiros passos, e um dos mais importantes, do processo de compreensão do genoma de uma espécie, uma vez que este é sequenciado total ou parcialmente.

Encontrar genes codificadores de proteínas em sequências genômicas de eucariotos através de métodos analíticos *in silico* é um trabalho computacional que possui diferentes abordagens de investigação. Os métodos existentes se enquadram em dois grandes grupos. O primeiro consiste em programas *ab initio*, que utilizam apenas sequências genômicas como base de dados. Exemplos disso são os programas GENSCAN [73], Augustus [74], HMMGene [75] e GENEID [76]. E o segundo é "expression based", isto é, prevê uma sequência gênica a partir de um gene homólogo que foi sequenciado anteriormente, ou de uma sequência protéica correspondente. Esta abordagem pode prever com precisão genes do genoma montado que são iguais ou muito semelhantes a genes que codificam transcritos homólogos já conhecidos [77].

Sistemas de previsão *de novo* de genes empregam modelos estatísticos para predizer as estruturas gênicas utilizando apenas as sequências de um ou mais genomas como base de

dados para análise. Outras sequências de cDNA ou dados de expressão não são necessárias, de modo que o métodos *de novo* podem prever novos genes a partir dos dados do genoma sequenciado. Tais metodologias ignoram, portanto, as sequências de cDNA que estão disponíveis. Por isto, esta abordagem tende a ser menos precisa do que aquelas baseadas em métodos em que as sequências de cDNA são usadas como referência (análise por homologia de sequência).

Estratégias para fornecer uma caracterização mais extensa do pool gênico empregam uma combinação dos dados genômicos de projetos de sequenciamento com dados obtidos a partir de sequenciamento do transcritomas. Neste caso, os resultados de sequenciamento usando RNA-seq e sequências genômicas montadas com a referência são analisados com o propósito de gerar transcrições de genoma para previsão de genes.

Programas alinhadores como Tophat [78] e PASA [79] possuem módulos que podem ser utilizados para a anotação do genoma e para modelar automaticamente estruturas gênicas. Além dos modelos de genes obtidos a partir de alinhamentos de transcritos de genoma eucariotos, e de preditores *ab initio* em separado, podem ser utilizados preditores de genes com abordagens para diferentes modelos de estruturas de genes, incluindo SNAP [80], 2011), GlimmerHMM [81] e Genemark_ES [82]. Uma vez disponibilizados diferentes conjuntos de dados resultantes de preditores de genes, um *pipeline* de ponderação pode ser usado para combinar os resultados destes preditores (*ab initio* e de homologia), tendo como resultado um conjunto consistente de genes anotados.

Estudos de espécies com baixo nível de informação molecular, como *Brachiaria*, podem valer-se das sequencias de ESTs em bancos de dados para a anotação gênica, metodologia que vem sendo utilizada para diferentes espécies [83]. Neste sentido, sequências de ESTs de *B. brizantha* usando o conjunto de *scaffolds* de *B.ruziziensis* como referência podem ser utilizadas para identificar o conjunto de genes ortólogos entre as duas espécies. Isto poderia ser o primeiro passo para incluir a *Brachiaria* nos estudos de ontologia de gramíneas. Esta metodologia foi testada no presente trabalho, como será descrito a seguir.

Elementos Repetitivos no Genoma

Elementos transponíveis são fragmentos de DNA que podem ser inseridos por movimentação física em novas localizações no genoma e, em alguns casos, podem fazer auto-cópias parciais ou integrais durante o processo de excisão de um local para inserção em outra região. Com o advento do sequenciamento em larga escala ficou claro que os elementos

transponíveis compreendem a maior parte do material genético de grande parte dos genomas eucariotos. Estes elementos representam pelo menos 45% do genoma humano [84] e de 50 a 90% do genoma de algumas plantas [85], como o milho e o pinheiro.

Em geral, para a análise de elementos repetitivos e elementos transponíveis no genoma, os programas disponíveis utilizam técnicas de verificação de similaridade entre sequências como ponto de partida. Alguns métodos para detecção de elementos transponíveis baseiam-se na prospecção de sequências consenso ou de elementos repetitivos, e podem considerar o conhecimento *a priori* sobre similaridades com sequências conhecidas. A estratégia mais comum é a detecção de pares de sequências similares em diferentes localidades do genoma do próprio organismo, seguida do agrupamento destes elementos para obter famílias de repetições. Em razão da não especificidade para elementos transponíveis, estes métodos invariavelmente encontram sequências oriundas de outros processos genéticos que também incluem repetições em tandem, duplicação segmental e satélites. Assim, o desafio no estudo é a distinção dos elementos transponíveis de outras classes de elementos e a identificação de diferentes famílias de elementos transponíveis.

A contribuição dos elementos transponíveis para a estrutura do genoma e evolução genômica, como também o impacto no sequenciamento, mapeamento e anotação, tem gerado um especial interesse no desenvolvimento de novos métodos computacionais para encontrar repetições. Além do modo único de replicação e abundância, os elementos transponíveis são entidades biológicas importantes devido ao seu papel na estrutura, tamanho e evolução dos genomas. A disponibilização dos bancos de dados de elementos transponíveis de referência para diferentes grupos de organismos [86] possibilita hoje uma adequada caracterização dos genomas em relação a estes componentes.

Dentre as dificuldades biológicas para o desenvolvimento de métodos mais eficientes de identificação e análise de elementos transponíveis, destacam-se aquelas ligadas à complexidade dos eventos biológicos associados aos elementos transponíveis. Como exemplo podem ser destacados o encadeamento de elementos transponíveis, a transcrição reversa incompleta e a existência de sequências de outras classes dentro da região do elemento transponível e, por fim, as similaridades entre famílias próximas [87].

Elementos transponíveis têm importância fundamental no que tange à composição e estrutura de genomas, pois são bastante numerosos na maior parte das espécies de eucariotos, principalmente naquelas com maior genoma. Inicialmente descritos em milho como elementos controladores [88], este tipo de estrutura tem sido bastante estudado em diversos organismos, e muitas funções e particularidades estruturais ainda estão por ser descritas.

O desenvolvimento de ferramentas genômicas para genotipagem de acessos de Brachiaria

Dentre os vários métodos que revelam polimorfismo de sequência de DNA conhecidos como marcadores moleculares, destacam-se os marcadores microssatélites (SSR – Single Sequence Repeats) e os marcadores SNP (*Single Nucleotide Polymorphism*) [89]. Marcadores microssatélites são definidos como repetições em tandem de pequenos motivos de DNA de 1 a 6 pb de comprimento que exibem variação no número de repetições num determinado loco [90-92]. Os microssatélites apresentam uma grande abundância genômica e multialelismo. O produto de amplificação das regiões microssatélites por PCR é utilizado como marcador molecular, revelando polimorfismo de comprimento em pares de bases de DNA.

O processo de desenvolvimento de marcadores microssatélite tradicionalmente envolve a construção de bibliotecas genômicas de pequenos fragmentos ou de bibliotecas enriquecidas para sequências hipervariáveis, seleção de colônias por hibridização, sequenciamento de clones selecionados, desenho de iniciadores para regiões que flanqueiam os elementos repetitivos, e verificação do nível de polimorfismo de cada marcador por PCR [89, 93].

Em braquiária, esforços de desenvolvimento de marcadores microssatélites a partir de bibliotecas enriquecidas foram realizados em *B. brizantha* e *B. humidicola*. Experimentos com transferibilidade de microssatélites entre espécies de braquiária com alguns marcadores desenvolvidos para *B. brizantha* e *B. humidicola* indicaram que poderiam ser gerados produtos de PCR em *B. ruziziensis* [94].

Outra abordagem na detecção e desenvolvimento de marcadores microssatélites é NGS do genoma, seguido de montagem *de novo* de sequências, e prospecção de regiões microssatélite. A abordagem NGS mais usada atualmente consiste na incorporação via PCR de uma base de nucleotídeo (A, C, G, T) em um *template* (sequência padrão) de DNA imobilizado em superfície sólida, usando bases modificadas que incluem um marcador para fluorescência e um terminador de reação. Após a captura de sinal fluorescente emitido pelo marcador, ambos, marcador e terminador, são removidos. O *template* de DNA pode então ser estendido com a incorporação da próxima base em um novo ciclo de sequenciamento. Esta abordagem foi recentemente aplicada na detecção, desenvolvimento e validação de locos microssatélite em *B. ruziziensis* (veja Capítulo 1 do presente trabalho).

Marcadores SNP podem ser definidos como marcadores que revelam uma substituição

de base na sequência de DNA entre amostras de indivíduos de uma mesma população [95]. Marcadores SNP normalmente possuem uma natureza bialélica, o que os torna menos informativos quando comparados aos microssatélites. Contudo, a abundância de SNPs no genoma compensa essa deficiência relativa de conteúdo informativo em relação a microssatélites [96]. Existem diversos métodos de detecção e desenvolvimento de SNPs. Inicialmente, os métodos mais utilizados baseavam-se no alinhamento de sequências obtidas de diversos indivíduos pela metodologia Sanger. Atualmente, as principais estratégias de obtenção de SNPs baseiam-se na avaliação de sequências de EST em bancos de dados e na seleção de SNPs a partir do sequenciamento NGS e montagem do genoma.

Os ensaios de genotipagem de SNPs incluem diferentes metodologias de detecção, como hibridização alelo-específica, extensão de *primer* e ligação de oligonucleotídeos. No presente trabalho, a ligação de oligonucleotídeos marcados com diferentes fluorescências para discriminar os alelos de um marcador SNP, empregando a tecnologia Infinium, foi testada em acessos e populações de *B. ruziziensis*. Deve ser destacada ainda o recente desenvolvimento da metodologia GBS (*Genotyping by Sequencing*), que combina o sequenciamento de alto desempenho com a descoberta e genotipagem simultânea de alelos em sítios SNP. O método ainda está em pleno desenvolvimento e apresenta algumas variações [97-99]. GBS permite genotipar milhares de marcadores SNPs, amplamente distribuídos ao longo do genoma, em um *pool* de amostras sequenciadas simultaneamente. A metodologia de GBS envolve: (a) redução de complexidade através de endonucleases; (b) ligação de adaptadores (c) NGS para sequenciamento em escala; (d) detecção e avaliação do polimorfismo revelado. Contudo, estudos recentes em arroz [38] indicam que apesar de ser uma tecnologia eficiente na detecção de polimorfismo SNP, a acurácia na detecção de genótipos (repetibilidade) ainda deve ser intensamente trabalhada para o uso em escala desta metodologia na genotipagem de plantas [38].

A disponibilidade de diferentes tipos de marcadores moleculares, como SNPs e microssatélites, capazes de detectar polimorfismo a um custo cada vez menor, aliada ao desenvolvimento de métodos estatísticos e softwares para a detecção de QTLs, permitiram a disseminação do uso de ferramentas moleculares no estudo de características quantitativas [100]. No presente trabalho, os primeiros passos para a obtenção de ferramentas genômicas de apoio a programa de melhoramento genético de *Brachiaria ruziziensis* são dados, com foco no desenvolvimento de marcadores microssatélites, indels e SNPs.

Sequenciamento em larga escala, marcadores moleculares e chips de DNA

Além de sua aplicação no sequenciamento e montagem de genomas completos, a tecnologia NGS tem também atraído muito interesse pela potencial identificação em larga escala de marcadores moleculares ao longo do genoma. A análise de sítios SNPs, conforme mencionado anteriormente, provê o desenvolvimento de uma importante ferramenta para mapeamento fino de regiões candidatas na determinação de haplótipos associados a características de interesse, para a seleção assistida por marcadores moleculares em programas de melhoramento genético, ou no processo de compreensão da base genética da diversidade fenotípica dentro e entre populações [101].

Milhares de marcadores SNPs, potencialmente informativos, podem ser utilizados no desenvolvimento de mapas genéticos de alta densidade, recurso essencial para a identificação de variações responsáveis por característica complexas ou QTLs. Os projetos de sequenciamento em larga escala oferecem a possibilidade de descoberta de SNPs a baixo custo, uma vez que as variações nas sequências podem ser verificadas computacionalmente, através da análise de bancos de dados de sequência [102].

As atuais tecnologias, antes mesmo de dispor da sequência completa de um determinado genoma, permitem realizar comparações entre sequências parciais para identificar polimorfismos, mutações e variações estruturais entre organismos. Essas ferramentas permitem a análise comparativa entre genomas em uma única execução experimental, possibilitando a cobertura necessária para a identificação correta de SNPs, além de variações estruturais, que podem envolver de kilobases a megabases, como inserções, deleções, variação no número de repetições e rearranjos. A dimensão destes experimentos pode ser evidenciada, por exemplo, em trabalho aplicado ao genoma de bovinos, no qual foi possível identificar 60.042 SNPs potenciais e predizer suas frequências alélicas, além de validar 92% de 23.357 SNPs selecionados ao longo do genoma [103].

Microarranjos de alta densidade para catalogação da variação de SNPs também têm sido usados em estudos de associação genótipo-fenótipo para identificar variações de sequência no genoma associadas a características de interesse. Esses microarranjos permitem genotipagem simultânea de milhares de SNPs em um grande número indivíduos, a um custo relativamente baixo. Os dois maiores produtores desses microarrajos, no momento, são Affymetrix Inc. (Santa Clara, CA) e Illumina Inc. (San Diego, CA). As plataformas oferecidas por essas empresas diferem substancialmente em termos de fabricação da sonda matriz, preparação de amostras e protocolo de hibridização. Atualmente, podem ser

genotipados até cerca de 1 milhão de SNPs por amostra, que também incluem sondas não polimórficas para avaliar a variação do número de cópias no genoma.

No caso da Illumina, a amostra de DNA utilizada para este ensaio é amplificada isotermicamente. A concentração de DNA exigida na amostra biológica é relativamente baixa, isto é, apenas 750 ng de DNA são suficientes para ensaio simultâneo (multiplex) de mais de milhares de SNPs. O produto amplificado é fragmentado por um processo enzimático controlado. Após a precipitação com álcool e re-suspensão do DNA, o *BeadChip*, como é chamado, é preparado para hibridização. As amostras de DNA fragmentadas e amplificadas são aneladas em locos específicos. Cada *bead* (semi-esfera de sílica) é utilizado para a detecção de um alelo SNP por loco. Depois da hibridização, a especificidade alélica é conferida por extensão enzimática de base. Os produtos são posteriormente corados por fluorescência e analisados computacionalmente. Esta técnica tem aumentado rapidamente tanto na densidade de SNP (de 3.000 a 1.000.000 SNPs) quanto no número de amostras processadas em paralelo (1, 2, 4, 8 ou 12 por BeadChip) ao longo dos últimos anos. No presente trabalho, o sequenciamento do genoma de *B. ruziziensis* a partir de tecnologia NGS e a consequente montagem parcial do genoma da espécie (veja Capítulo 2) permitiu a seleção de milhares de SNPs. Estes dados serão usados na confecção de um chip de detecção de polimorfismo de DNA, que contribuirá para inaugurar uma nova etapa na genotipagem de acessos de *B .ruziziensis*, com impacto na conservação de germoplasma e melhoramento genético da espécie.

.

Sequenciamento e montagem de genomas de cloroplastos por NGS e desenvolvimento de marcadores indel para identificação de espécies de braquiária

Embora o gênero tenha sido reconhecido no século XIX, a classificação botânica de braquiária ainda gera controvérsias, conforme mencionado anteriormente. Os limites precisos para diferenciação de *Brachiaria* de gêneros próximos como *Urochloa*, *Eriochloa* e *Panicum* têm provocado discussão [3, 11, 32]. Há, inclusive, uma proposta de transferência de algumas das espécies mais importantes de *Brachiaria* para o gênero *Urochloa* [11], que poderia ainda incluir *Panicum maximum*. Uma análise filogenética recente concluiu que *Brachiaria* e *Urochloa* formam um grupo monofilético (junto com *Eriochloa* e *Melinis*) [3]. A análise da variação genômica pode contribuir para a a melhor compreensão da filogenia das espécies próximas de *Brachiaria*.

A variação genômica pode ser focada no genoma nuclear, cloroplástico ou mitocondrial. Quando são analisadas as variações de tamanho, organização e sequência do genoma nuclear, cloroplástico e mitocondrial da plantas, o cpDNA é considerado evolutivamente mais conservado que os demais. A variação em tamanho, por exemplo, é relativamente pequena, visto que o maior cpDNA já registrado não excede mais do que o dobro do tamanho do menor genoma cloroplástico até agora identificado. Entre as milhares de espécies de plantas analisadas, o tamanho do cpDNA varia apenas de 120 a 210 Kbp [14]. Em contrapartida, o genoma nuclear de plantas apresenta variações de uma ou mais ordens de magnitude em tamanho. Somente entre as espécies agrícolas de uma mesma família, como as gramíneas, observa-se uma variação de 30x de tamanho de genoma entre o arroz e o trigo. Por sua vez, o genoma mitocondrial é, via de regra, substancialmente maior (e mais variável) do que o genoma cloroplástico. Grandes variações de tamanho na estrutura do cpDNA, como inserções e deleções, também são raras, assim como transposições e inversões, embora sejam fenômenos comuns no genoma nuclear.

O genoma cloroplástico, portanto, apresenta um conjunto de características que o qualificam para análises filogenéticas, visto que o genoma nuclear e mitocondrial são mais dinâmicos na sua diversificação, especialmente porque submetidos ao processo de recombinação (*crossing-over*), que não ocorre no genoma cloroplástico. Processos evolutivos comuns no genoma nuclear como duplicação ou deleção gênica, incluindo a maciça presença de famílias gênicas, praticamente não são presentes no cpDNA. Além disso, os demais genomas são maiores (em várias ordens de magnitude) e mais complexos, e apresentam grande quantidade de sequências repetitivas, duplicações e inversões - especialmente o genoma nuclear - o que dificulta a interpretação taxonômica.

A documentação da variação da sequência de cloroplastos tem sido uma ferramenta essencial em estudos evolutivos e de populações de plantas por várias décadas. Com um tamanho médio de 120-160 kb, e contendo ~130 genes, genomas de cloroplastos são suficientemente grandes e complexos a ponto de incluir mutações estruturais e sítios de diferenciação em nível de população, possibilitando a avaliação de divergências evolutivas entre espécies.

Na maioria das plantas terrestres, genomas de cloroplasto consistem em um único cromossomo circular, com uma estrutura quadripartida, que inclui uma região grande de cópia única (LSC) e uma região pequena de cópia simples (SSC), separadas por duas cópias de repetições invertidas (IR). O conteúdo genético, ordem e organização dos genomas de cloroplasto geralmente são altamente conservados e a herança genética é principalmente

materna. Tal modo de herança uniparental faz dos genomas de cloroplasto valiosas estruturas para estudos de genética e de filogenia [104].

Uma característica importante do genoma de cloroplasto é o seu elevado grau de conservação das sequências. Seleção natural intensa, agindo em maquinaria fotossintética, impõe restrições claras sobre as taxas de mutação de nucleotídeos. Devido a estas restrições, as alterações estruturais nas regiões não codificadoras são muitas vezes utilizados para estudar diferenciação de populações de plantas, enquanto que as sequências codificadoras do genoma cloroplástico, bastante conservadas, têm sido usadas com sucesso para resolver relações filogenéticas entre organismos. Devido aos limites severos impostos na divergência das sequências de cloroplasto, comprimentos significativos da sequência de DNA de cloroplasto são muitas vezes necessários para detectar estatisticamente a diferenciação da população ou resolução filogenética [105].

Sequências do genoma de cloroplasto contêm, portanto, algumas regiões que são variáveis entre espécies. Essas regiões têm sido consideradas extensivamente na seleção dos locos adequados para distinguir espécies estreitamente relacionadas ou gêneros em análises filogenéticas [106]. Conforme mencionado anteriormente, a distinção de acessos de espécies de *Brachiaria*, especialmente *B. ruziziensis, B. decumbens* e *B. brizantha*, é muito difícil morfologicamente em alguns estádios de desenvolvimento e passível de erro. A análise do genoma cloroplástico destas espécies pode auxiliar na classificação de acessos de braquiária. Este passo é fundamental, por exemplo, para facilitar a coleta de acessos de braquiária fora da época de florescimento, fomentando a formação de bancos de germoplasma para a conservação em longo prazo de estoques genéticos e para uso no programa de melhoramento.

Uma forma de utilização do genoma cloroplasto em análise filogenética e na diferenciação de espécies é o emprego de regiões "universais" deste genoma, consideradas *hot spots* de polimorfismo de DNA. Estas regiões são conhecidas como regiões *barcoding* ou "sistema de barras codificadas" de DNA cloroplástico para discriminação de espécies [107]. Isto porque os haplótipos observados nestas regiões seriam, teoricamente, típicos de cada espécie e, portanto, poderiam ser usados na sua discriminação. *DNA barcoding*, portanto, engloba um conjunto de metodologias que utiliza marcadores de regiões específicas ("universais") de DNA nuclear, cloroplástico ou mitocondrial (no caso de animais) de um organismo para identificá-lo como pertencente a uma espécie particular e diferenciá-lo de espécies afins. A análise de *DNA barcoding* difere de estudos de filogenia molecular ao focar na identificação da espécie de uma amostra desconhecida, possibilitando a sua classificação, e não na determinação de seu vínculo genético com espécies próximas e distantes [108].

Apesar de *barcoding* ser por vezes usado em um esforço para identificar espécies desconhecidas ou avaliar se as espécies devem ser combinadas ou separadas, a sua aplicação para esse fim tem gerado controvérsia [109]. Alguns pesquisadores argumentam que em certas situações a caracterização *barcoding* não fornece informações confiáveis ao nível de espécie, mas que ainda pode ter mérito para um nível superior de classificação [108].

No caso das espécies de *Brachiaria*, para diferenciação por *barcoding*, torna-se necessária a existência de uma porção no DNA que possa ser polimórfico o suficiente para distinguir diferentes espécies do gênero. Isto poderia ser encontrado, por exemplo, em regiões conservadas localizados no genoma cloroplástico, como as regiões gênicas *rbcL* e *matK*, que são habitualmente recomendadas para esta propósito [107]. Para inferência filogenética, a utilização de um maior número de regiões gênicas ou de polimorfismo de sequência de DNA é recomendável para minimizar o ruído da análise focada em um ou poucos genes, devido à heterogeneidade evolutiva de genes ou partes de um gene. Felizmente, existem no momento muitas sequências completas do genoma de cloroplasto disponíveis, incluindo aquelas de um mesmo gênero. Esse banco de dados permite a identificação da maioria das regiões variáveis entre ou dentro de espécies com base na sequência do DNA cloroplástico [106].

Através da combinação de clones de cromossomos artificiais de bactérias (BAC) e sequenciamento NGS, o genoma de cloroplasto (cpDNA) de plantas pode ser sequenciado com precisão, eficiência e economia [110]. A busca por sítios polimórficos espécie-específicos pode ser feita diretamente a partir do sequenciamento e alinhamento dos genomas de cloroplastos montados desta forma. No presente trabalho, o sequenciamento NGS do genoma cloroplástico de *B. ruziziensis, B. decumbens, B. brizantha* e *B. humidicola* foi realizado, processado e analisado para a identificação de regiões específicas capazes de possibilitar a identificação e discriminação de acessos destas diferentes espécies (veja Capítulo 3).

Referências

1. Rosengurtt, B., B.A. de Maffei, and P.I. De Artucio, *Gramíneas uruguayas*. Vol. 5. 1970: Universidad de la República, Departamento de Publicaciones.
2. T, S., *Chave para identificação de Brachiaria.* Agroceres, 1977. **V**: p. 4-5.
3. Torres González, A.M. and C.M. Morton, *Molecular and morphological phylogenetic analysis of Brachiaria and Urochloa (Poaceae).* Molecular Phylogenetics and Evolution, 2005. **37**(1): p. 36-44.
4. Trinius, *Panicearum genera / retractavit speciebusque compluribus illustravit C.B. Trinius*. 1834, St.-Pétersbourg :: Impr. de l'Académie impériale des sciences.
5. Renvoize, S., et al., *Morfología, taxonomía y distribución natural de Brachiaria (Trin.) Griseb.* Brachiaria: Biología, Agronomía y Mejoramiento. CIAT. Cali, Colombia, 1998: p. 1-17.
6. Bogdan, A.V., *Tropical pasture and fodder plants*. 1977, London: Longman. xiii + 475 pp.
7. Alcantara, P.B. and G. Bufarah, *Plantas forrageiras: gramíneas & leguminosas*. 1986: Nobel.
8. Soares Filho, C., F. Monteiro, and M. Corsi, *Recuperação de pastagens degradadas de Brachiaria decumbens. 1. Efeito de diferentes tratamentos de fertilização e manejo.* Pasturas Tropicales, 1992. **14**(2): p. 1-6.
9. Assis, G.M.L.d., et al., *Discriminação de espécies de Brachiaria baseada em diferentes grupos de caracteres Morfológicos.* Revista Brasileira de Zootecnia, 2003. **32**: p. 576-584.
10. Maass, B., et al., *Identificación y nomenclatura de las especies de Brachiaria.* Brachiaria: biología, agronomía y mejoramiento, 1998.
11. Webster, R.D., *Australian Paniceae (Poaceae).* 1987: J. Cramer.
12. Morrone, O. and F.O. Zuloaga, *Sinopsis del género Urochloa Poaceae: Panicoideae: Paniceae para México y América Central.* 1993. **v. 32, no. 1/4, p. 59-75**.
13. Loch, D., *Brachiaria decumbens (signal grass): a review with particular reference to Australia.* Trop. Grasslands, 1977. **11**(2): p. 141-157.
14. Clayton, W.D., Vorontsova, M.S., Harman, K.T. and Williamson, H., *GrassBase - The Online World Grass Flora.* 2006.
15. Sánchez-Ken, J.G., *A synopsis of Digitaria (Paniceae, Panicoideae, Poaceae) in Mexico, including the new species Digitaria michoacanensis.* Acta botánica mexicana, 2012: p. 127-149.
16. A.V., B., *The selection of tropical ley grasses in Kenia: general considerations and methods.* East. Afr. Agrc. J, 1959. **24**: p. 206-217.
17. Brown, W.V. and W.H. Emery, *Apomixis in the Gramineae: Panicoideae.* American Journal of Botany, 1958: p. 253-263.
18. Pritchard, A.J., *Apomixis in Brachiaria decumbens Stapf.* Aust. Inst. agric. Sci, 1967. **33**(4): p. 264-265.
19. Valle, C.B.d. and Y.H. Savidan, *Genetics, cytogenetics and reproductive biology of Brachiaria.* Brachiaria: biology, agronomy, and improvement, 1996.
20. Penteado, M.d.O., et al., *Determinação de ploidia e avaliação da quantidade de DNA total em diferentes espécies do gênero Brachiaria*. 2000: Embrapa Gado de Corte.
21. Ishigaki, G., et al., *Estimation of genome size in Brachiaria species.* Grassland Science, 2010. **56**(4): p. 240-242.
22. Ferguson, J. and L. Crowder, *Cytology and breeding behavior of Brachiaria*

ruziziensis Germain et Evrard. Crop Science, 1974. **14**(6): p. 893-895.
23. Sotomayor-Ríos, A., S. Schank, and R. Woodbury, *Cytology and taxonomic description of two Brachiaria [spp.](Congograss and Tanner-grass).* Journal of Agriculture of the University of Puerto Rico, 1970. **54**(2): p. 390-400.
24. McGregor Jr, J.T., R.J. Smith Jr, and R.E. Talbert, *Broadleaf signalgrass (Brachiaria platyphylla) duration of interference in rice (Oryza sativa).* Weed Science, 1988: p. 747-750.
25. Seiffert, N.F., *Gramíneas forrageiras do gênero Brachiaria*. 1980: EMBRAPA, Centro Nacional de Pesquisa de Gado de Corte.
26. Holmann, F., et al., *Impact of the adoption of Brachiaria grasses: Central America and Mexico.* Livestock Research for Rural Development, 2004. **16**(12): p. 1-9.
27. Boddey, R.M., et al., *Nitrogen cycling in Brachiaria pastures: the key to understanding the process of pasture decline.* Agriculture, Ecosystems & Environment, 2004. **103**(2): p. 389-403.
28. Lima, E.d.V., et al., *Mistura de sementes de Brachiaria brizantha com fertilizante NPK.* Ciência Rural, 2010. **40**: p. 441-444.
29. BARCELLOS, A.d.O.A., R.P. de; KARIA, C.T.; VILELA, L., *Potencial e uso de leguminosas forrageiras dos gêneros Stylosanthes, Arachis e Leucaena.* . SIMPÓSIO SOBRE MANEJO DA PASTAGEM, 17 2001. **17**.
30. Barbosa, J.D., et al., *Fotossensibilização hepatógena em eqüinos pela ingestão de Brachiaria humidicola (Gramineae) no Estado do Pará.* Pesq. Vet. Bras, 2006. **26**(3): p. 147-153.
31. Karia, C.T., J.B. Duarte, and A.d. Araújo, *Desenvolvimento de cultivares do gênero Brachiaria (trin.) Griseb. no Brasil*. 2006: Embrapa Cerrados.
32. Kumble, V., et al., *Brachiaria: biology, agronomy, and improvement*. 1996: CIAT.
33. Serrão, E.A.S. and M.S. Neto, *Informacões sôbre duas espécies de gramíneas forrageiras do gênero Brachiarìa na Amazônia: B. decumbens Stapf e B. ruziziensis Germain et Everard*. 1971: Instituto de Pesquisas e Experimentação Agropecuárias do Norte.
34. Valério, J.R.O., M. C. M. , *Parasitismo de ovos de cigarrinhas-das-pastagens (Homoptera: Cercopidae) pelo microhimenóptero Anagrus urichi Pickles (Hymenoptera: Mymaridae) na região de Campo Grande, MS.*
Neotropical Entomology 2005. **34**: p. 137-138.
35. Clayton, W., K. Harman, and H. Williamson, *onwards. World grass species: descriptions, identification, and information retrieval*. 2002.
36. SOUZA SOBRINHO, F.d.L., F. J. da S.; KOPP, M. M.; PEREIRA, A. V.; SOUZA, F. F. DE, *Melhoramento de gramíneas forrageiras na Embrapa Gado de Leite.* In: SIMPÓSIO E CONGRESSO DE FORRAGICULTURA E PASTAGENS, 7., 2009.
37. Miles, J.W. and C.B. do Valle, *Brachiaria: Biología, agronomía y mejoramiento*. 1998: CIAT.
38. Silva, P.I., et al., *Development and validation of microsatellite markers for Brachiaria ruziziensis obtained by partial genome assembly of Illumina single-end reads.* BMC Genomics, 2013. **14**(1): p. 17.
39. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.
40. Sanger, F., et al., *Nucleotide sequence of bacteriophage [phi]X174 DNA*. Nature, 1977. **265**(5596): p. 687-695.
41. Hutchison, C.A., 3rd, *DNA sequencing: bench to bedside and beyond.* Nucleic Acids Res, 2007. **35**(18): p. 6227-37.
42. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre*

reactors. Nature, 2005. **437**(7057): p. 376-80.
43. Metzker, M.L., *Sequencing technologies—the next generation.* Nature Reviews Genetics, 2009. **11**(1): p. 31-46.
44. Wold, B. and R.M. Myers, *Sequence census methods for functional genomics.* Nat Methods, 2008. **5**(1): p. 19-21.
45. Thompson, J.F. and K.E. Steinmann, *Single molecule sequencing with a HeliScope genetic analysis system.* Current Protocols in Molecular Biology, 2010: p. 7.10. 1-7.10. 14.
46. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-380.
47. Shendure, J., et al., *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. **309**(5741): p. 1728-1732.
48. Huse, S.M., et al., *Accuracy and quality of massively parallel DNA pyrosequencing.* Genome Biol, 2007. **8**(7): p. R143.
49. Pearson, B.M., et al., *The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828).* J Bacteriol, 2007. **189**(22): p. 8402-3.
50. Barbazuk, W.B., et al., *SNP discovery via 454 transcriptome sequencing.* Plant J, 2007. **51**(5): p. 910-8.
51. Huber, J.A., et al., *Microbial population structures in the deep marine biosphere.* Science, 2007. **318**(5847): p. 97-100.
52. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.
53. Liu, S., et al., *<italic>De Novo</italic> Transcriptome Assembly in Chili Pepper (<italic>Capsicum frutescens</italic>) to Identify Genes Involved in the Biosynthesis of Capsaicinoids.* PLoS ONE, 2013. **8**(1): p. e48156.
54. Lu, B., et al., *Effective driving force applied on DNA inside a solid-state nanopore.* Phys Rev E Stat Nonlin Soft Matter Phys, 2012. **86**(1 Pt 1): p. 011921.
55. Xu, M., D. Fujita, and N. Hanagata, *Perspectives and Challenges of Emerging Single-Molecule DNA Sequencing Technologies.* Small, 2009. **5**(23): p. 2638-2649.
56. Chaisson, M.J., D. Brinza, and P.A. Pevzner, *De novo fragment assembly with short mate-paired reads: Does the read length matter?* Genome Res, 2009. **19**(2): p. 336-46.
57. Chevreux, B., et al., *Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.* Genome Res, 2004. **14**(6): p. 1147-59.
58. Dohm, J.C., et al., *SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing.* Genome Research, 2007. **17**(11): p. 000.
59. Zhang, W., et al., *A Practical Comparison of <italic>De Novo</italic> Genome Assembly Software Tools for Next-Generation Sequencing Technologies.* PLoS ONE, 2011. **6**(3): p. e17915.
60. Koonin, E.V., *Computational genomics.* Curr Biol, 2001. **11**(5): p. R155-8.
61. Baker, M., *De novo genome assembly: what every biologist should know.* Nat Meth, 2012. **9**(4): p. 333-337.
62. Salzberg, S.L., et al., *GAGE: A critical evaluation of genome assemblies and assembly algorithms.* Genome Research, 2012. **22**(3): p. 557-567.
63. Compeau, P.E.C., P.A. Pevzner, and G. Tesler, *How to apply de Bruijn graphs to genome assembly.* Nat Biotech, 2011. **29**(11): p. 987-991.
64. Peng, Y., et al., *IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler*, in *Research in Computational Molecular Biology*, B. Berger, Editor. 2010, Springer Berlin Heidelberg. p. 426-440.

65. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
66. Warren, R.L., et al., *Assembling millions of short DNA sequences using SSAKE.* Bioinformatics, 2007. **23**(4): p. 500-501.
67. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res, 2010. **20**(2): p. 265-72.
68. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
69. Aury, J.-M., et al., *High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies.* BMC Genomics, 2008. **9**(1): p. 603.
70. Garg, R., et al., *De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification.* DNA Research, 2011. **18**(1): p. 53-63.
71. Liu, Y., et al., *Complete chloroplast genome sequences of Mongolia medicine Artemisia frigida and phylogenetic relationships with other plants.* PLoS One, 2013. **8**(2): p. e57533.
72. Li, R., et al., *The sequence and de novo assembly of the giant panda genome.* Nature, 2010. **463**(7279): p. 311-317.
73. Burge, C. and S. Karlin, *Prediction of complete gene structures in human genomic DNA.* J Mol Biol, 1997. **268**(1): p. 78-94.
74. Stanke, M. and S. Waack, *Gene prediction with a hidden Markov model and a new intron submodel.* Bioinformatics, 2003. **19**(suppl 2): p. ii215-ii225.
75. Krogh, A., *Two methods for improving performance of an HMM and their application for gene finding.* Center for Biological Sequence Analysis. Phone, 1997. **45**: p. 4525.
76. Blanco, E., G. Parra, and R. Guigó, *Using geneid to identify genes.* Current protocols in bioinformatics, 2007: p. 4.3. 1-4.3. 28.
77. Wei, C. and M.R. Brent, *Using ESTs to improve the accuracy of de novo gene prediction.* BMC Bioinformatics, 2006. **7**: p. 327.
78. Kim, D. and S.L. Salzberg, *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome Biol, 2011. **12**(8): p. R72.
79. Haas, B.J., et al., *Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.* Nucleic acids research, 2003. **31**(19): p. 5654-5666.
80. Zaharia, M., et al., *Faster and more accurate sequence alignment with SNAP.* arXiv preprint arXiv:1111.5572, 2011.
81. Delcher, A.L., et al., *Identifying bacterial genes and endosymbiont DNA with Glimmer.* Bioinformatics, 2007. **23**(6): p. 673-679.
82. Ter-Hovhannisyan, V., et al., *Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training.* Genome Res, 2008. **18**(12): p. 1979-90.
83. Morozova, O., M. Hirst, and M.A. Marra, *Applications of new sequencing technologies for transcriptome analysis.* Annual review of genomics and human genetics, 2009. **10**: p. 135-151.
84. Mills, R.E., et al., *Which transposable elements are active in the human genome?* TRENDS in Genetics, 2007. **23**(4): p. 183-191.
85. Meyers, B.C., S.V. Tingey, and M. Morgante, *Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome.* Genome Res, 2001. **11**(10): p. 1660-76.
86. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.
87. Bergman, C.M. and H. Quesneville, *Discovering and detecting transposable elements in genome sequences.* Briefings in Bioinformatics, 2007. **8**(6): p. 382-392.
88. McClintock, B., *The origin and behavior of mutable loci in maize.* Proceedings of the

National Academy of Sciences, 1950. **36**(6): p. 344-355.
89. Ferreira, M.E. and D. Grattapaglia, *Introdução ao uso de marcadores RAPD e RFLP em análise genética*. 1995: Embrapa-Cenargen.
90. Litt, M. and J.A. Luty, *A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene.* Am J Hum Genet, 1989. **44**(3): p. 397-401.
91. Tautz, D., *Hypervariabflity of simple sequences as a general source for polymorphic DNA markers.* Nucleic acids research, 1989. **17**(16): p. 6463-6471.
92. Weber, J.L. and P.E. May, *Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction.* Am J Hum Genet, 1989. **44**(3): p. 388-96.
93. Ritschel, P.S., et al., *Development of microsatellite markers from an enriched genomic library for genetic analysis of melon (Cucumis melo L.).* BMC Plant Biology, 2004. **4**(1): p. 9.
94. Cançado, L.J.C., *Caracterização da diversidade genética molecular em germoplasma de Brachiaria spp*. 2009.
95. Risch, N. and K. Merikangas, *The Future of Genetic Studies of Complex Human Diseases.* Science, 1996. **273**(5281): p. 1516-1517.
96. Kruglyak, L., *The use of a genetic map of biallelic markers in linkage studies.* Nat Genet, 1997. **17**(1): p. 21-4.
97. Baird, N.A., et al., *Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers.* PLoS ONE, 2008. **3**(10): p. e3376.
98. Huang, X., et al., *High-throughput genotyping by whole-genome resequencing.* Genome Res, 2009. **19**(6): p. 1068-76.
99. Elshire, R.J., et al., *A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.* PloS one, 2011. **6**(5): p. e19379.
100. Bernardo, R., *Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.* Crop Sci., 2008. **48**(5): p. 1649-1664.
101. Panitz, F., et al., *SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation.* Bioinformatics, 2007. **23**(13): p. i387-91.
102. Parsons, A.B., et al., *Exploring the Mode-of-Action of Bioactive Compounds by Chemical-Genetic Profiling in Yeast.* Cell, 2006. **126**(3): p. 611-625.
103. Van Tassell, C., et al., *SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries.* Nat Methods, 2008. **5**(3): p. 247 - 252.
104. Young, H.A., et al., *Chloroplast genome variation in upland and lowland switchgrass.* PLoS One, 2011. **6**(8): p. e23980.
105. Cronn, R., et al., *Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology.* Nucleic acids research, 2008. **36**(19): p. e122-e122.
106. Dong, W., et al., *Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding.* PLoS ONE, 2012. **7**(4): p. e35071.
107. Group, C.P.W., et al., *A DNA barcode for land plants.* Proceedings of the National Academy of Sciences, 2009. **106**(31): p. 12794-12797.

108. Kress, W.J., et al., *Use of DNA barcodes to identify flowering plants.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(23): p. 8369-8374.
109. Koch, H., *Combining morphology and DNA barcoding resolves the taxonomy of western Malagasy Liotrigona Moure, 1961 (Hymenoptera: Apidae: Meliponini).* African Invertebrates, 2010. **51**(2): p. 413-421.
110. Pan, I.C., et al., *Complete Chloroplast Genome Sequence of an Orchid Model Plant Candidate: <italic>Erycina pusilla</italic> Apply in Tropical <italic>Oncidium</italic> Breeding*. PLoS ONE, 2012. **7**(4): p. e34738.

IV. Justificativa

A alta vulnerabilidade genética dos pastos de braquiária no Brasil é causada pela baixa diversidade genética das variedades plantadas. São pouquíssimas variedades de braquiária disponibilizadas para o produtor, que cobrem áreas muito extensas das terras cultivadas, atingindo cerca de 80 milhões de hectares. A braquiária é a forrageira com maior área plantada no Brasil e, portanto, a mais importante neste segmento do agronegócio. Não há dúvida de que aumentar a diversidade genética dos plantios de braquiária é estratégico para a pecuária brasileira.

A escolha da espécie *B. ruziziensis* como foco do trabalho justifica-se pelos seguintes aspectos: (a) possui biologia diplóide (2n=2x=18); (b) modo de reprodução é sexual; (c) há possibilidade de emprego imediato de métodos convencionais de melhoramento genético no desenvolvimento de novas cultivares; (d) possui genoma relativamente pequeno (~600 Mpb – Ishigaki et al., 2010), similar ao de outras espécies modelo como arroz (420 Mpb) ou sorgo (700 Mpb), o que facilita iniciativas de análise do genoma e de desenvolvimento de ferramentas genômicas para apoio ao programa de melhoramento genético; (e) apresenta boa qualidade forrageira, reconhecidamente a mais alta entre espécies de braquiária (Sobrinho et al., 2009); (f) possui excelente adaptabilidade a sistemas de produção integrada com lavoura pasto e floresta, tanto para alimentação animal (verde ou palhada), ou como cobertura de solo para plantio direto; (g) é passível de cruzamento, após tetraploidização, com outras espécies de braquiária de grande interesse para o agronegócio, como *B. decumbens* e *B. brizantha,* facilitando a introgressão de genes de uma espécie para outra e possibilitando a diversificação genética em nível tetraplóide através de cruzamentos interespecíficos.

O desenvolvimento de soluções para o problema da vulnerabilidade genética de braquiária através do melhoramento genético requer o desenvolvimento de métodos e tecnologias de genética molecular e de genômica que possam apoiar a dinamização da oferta de novas cultivares de braquiária para o mercado brasileiro. *B. ruziziensis*, pelas características listadas acima, apresenta-se como um excelente modelo entre as espécies de braquiária para o desenvolvimento de ferramentas genômicas. No momento, não há nenhuma ferramenta genômica (ex. marcadores moleculares ou painéis de genotipagem em escala) disponível para análise genética de *B. ruziziensis*. A análise *in silico* do genoma nuclear e cloroplástico de *B. ruziziensis*, a partir de segmentos de leitura NGS (*Next Generation Sequencing*) do genoma nuclear e cloroplástico obtidos no presente estudo, visa contribuir

para o desenvolvimento, validação e aplicação exitosa das ferramentas genômicas para apoio à conservação de germoplasma e ao melhoramento genético da espécie.

V. Objetivo geral

Este estudo tem como foco o desenvolvimento e uso de ferramentas de bioinformática aplicadas à análise de grandes volumes de dados de sequenciamento para identificar e selecionar variações específicas de sequência de DNA, como polimorfismos de único nucleotídeo (SNP - *Single Nucleotide Polymorphism*), marcadores microssatélites (SSR – *Single Sequences Repeats*) e (indels - *Insertions/Deletions*), visando o seu emprego em programas de conservação de germoplasma e de melhoramento genético de *Brachiaria ruziziensis*. Além disto, pretende valer-se das análises *in silico* com base no genoma de espécies conhecidas como modelo para estudo (ex. arroz), para a caracterização do genoma de *Brachiaria ruziziensis,* uma espécie órfã de informação genômica.

Objetivos específicos

a. Sequenciar, montar *de novo*, analisar e caracterizar o genoma estrutural de *Brachiaria ruziziensis*, com ênfase no conhecimento da composição de elementos transponíveis, bem como do espaço gênico, em comparação com outras espécies;
b. Desenvolver marcadores microssatélites para uso em análise genética e no programa de melhoramento de *B. ruziziensis* através de sequenciamento de alto desempenho (NGS – *Next Generation Sequencing*) do genoma nuclear de braquiária.
c. Sequenciar, montar *de novo*, analisar e caracterizar o genoma cloroplástico das quatro principais espécies de braquiária no Brasil (*B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola*). Desenvolver e validar marcadores espécie-específicos baseados inserções/deleções do DNA cloroplástico para a identificação de acessos destas espécies.
d. Desenvolver marcadores SNPs para uso em análise genética e no programa de melhoramento de *B. ruziziensis* através de NGS.

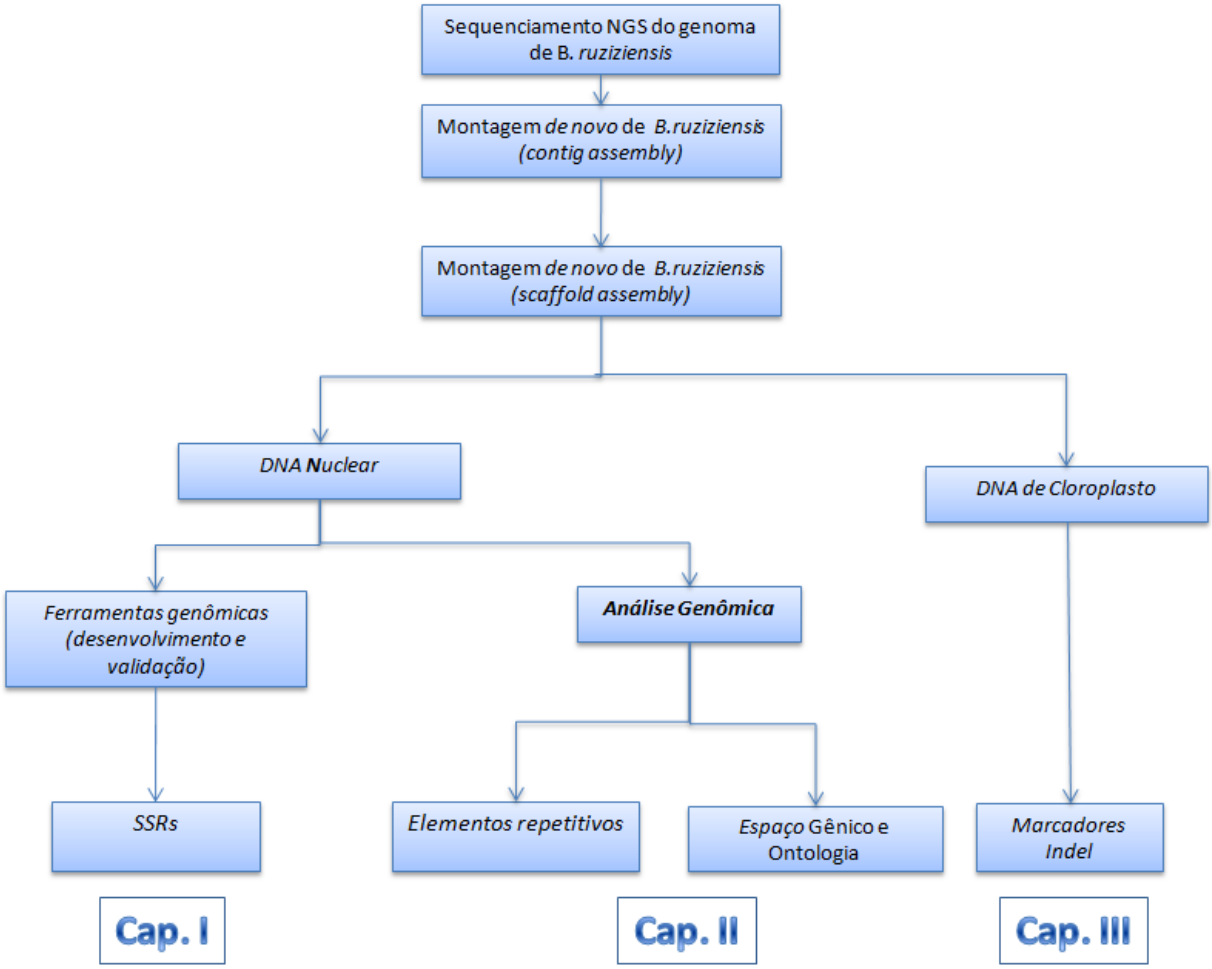
VI Plano de Tese

CAPÍTULO 1: Desenvolvimento, análise e validação de marcadores microssatélites de *B. ruziziensis*.

CAPÍTULO 2: Sequenciamento, montagem *de novo* e análise do genoma de *Brachiaria ruziziensis*

CAPÍTULO 3: Montagem e caracterização do genoma de cloroplasto de quatro espécies de *Brachiaria* e desenvolvimento de marcadores *indel* para diferenciação de espécies do gênero.

VII. Fluxograma



VIII. CAPÍTULO 1

Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads

Publicado como:
Silva PI, Martins AM, Gouvea EG, Pessoa-Filho M, Ferreira ME. 2013 Development and validation of microsatellite markers for Brachiaria ruziziensis obtained by partial genome assembly of Illumina single-end reads. BMC Genomics. 16;14:17. doi: 10.1186/1471-2164-14-17

Silva et al. BMC Genomics 2013, 14:17
http://www.biomedcentral.com/1471-2164/14/17

BMC Genomics

RESEARCH ARTICLE Open Access

Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads

Pedro IT Silva[1,2,4†], Alexandre M Martins[1,2†], Ediene G Gouvea[1], Marco Pessoa-Filho[3] and Márcio E Ferreira[1*]

Abstract

Background: *Brachiaria ruziziensis* is one of the most important forage species planted in the tropics. The application of genomic tools to aid the selection of superior genotypes can provide support to *B. ruziziensis* breeding programs. However, there is a complete lack of information about the *B. ruziziensis* genome. Also, the availability of genomic tools, such as molecular markers, to support *B. ruziziensis* breeding programs is rather limited. Recently, next-generation sequencing technologies have been applied to generate sequence data for the identification of microsatellite regions and primer design. In this study, we present a first validated set of SSR markers for *Brachiaria ruziziensis*, selected from a *de novo* partial genome assembly of single-end Illumina reads.

Results: A total of 85,567 perfect microsatellite loci were detected in contigs with a minimum 10X coverage. We selected a set of 500 microsatellite loci identified in contigs with minimum 100X coverage for primer design and synthesis, and tested a subset of 269 primer pairs, 198 of which were polymorphic on 11 representative *B. ruziziensis* accessions. Descriptive statistics for these primer pairs are presented, as well as estimates of marker transferability to other relevant brachiaria species. Finally, a set of 11 multiplex panels containing the 30 most informative markers was validated and proposed for *B. ruziziensis* genetic analysis.

Conclusions: We show that the detection and development of microsatellite markers from genome assembled Illumina single-end DNA sequences is highly efficient. The developed markers are readily suitable for genetic analysis and marker assisted selection of *Brachiaria ruziziensis*. The use of this approach for microsatellite marker development is promising for species with limited genomic information, whose breeding programs would benefit from the use of genomic tools. To our knowledge, this is the first set of microsatellite markers developed for this important species.

Background

The area planted with forage crops in the tropics extends for hundreds of millions of hectares. In Brazil alone, the forage cropped land exceeds 100 M ha [1], where four brachiaria species (*B. brizantha, B. decumbens, B. ruziziensis* and *B. humidicola*) cover 85% of the cultivated pastures [2]. Only a few apomictic brachiaria clones occupy tens of millions of hectares in the country [3], what represents a high risk of genetic vulnerability for forage production. This risk could be reduced with the increased use of genetic diversity conserved in germplasm banks in order to generate recombinant genotypes in breeding programs. The development and adoption of new brachiaria cultivars with a broad genetic base is crucial for the diversification of forage pasture in the tropics. The development of new cultivars must be a dynamic process, providing the pasture production sector with increasing genetic diversity.

Among the four brachiaria species most cultivated in Brazil, ruzigrass (*Brachiaria ruziziensis*, 2n=2x=18) stands out as a diploid species with sexual reproduction. Polyploid

* Correspondence: marcio.ferreira@embrapa.br
†Equal contributors
[1]Embrapa Recursos Genéticos e Biotecnologia, Genetics Lab, PO Box 02372, Brasília CEP 70770-917Distrito Federal, Brazil
Full list of author information is available at the end of the article

CAPÍTULO 1

Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads

Pedro IT Silva[1,2,4,†]
Email: tanno.pedro@gmail.com

Alexandre M Martins[1,2,†]
Email: mm.alexandre@gmail.com

Ediene G Gouvea[1]
Email: edienegouvea@gmail.com

Marco Pessoa-Filho[3]
Email: marco.pessoa@embrapa.br

Márcio E Ferreira[1*]
[*] Corresponding author
Email: marcio.ferreira@embrapa.br

[1] Embrapa Recursos Genéticos e Biotecnologia, Genetics Lab, PO Box 02372, Brasília CEP 70770-917, Distrito Federal, Brazil

[2] Departamento de Biologia Celular, IB - Universidade de Brasília (UnB) Campus Universitário Darcy Ribeiro, Asa Norte, Brasília CEP 70910-900, Distrito Federal, Brazil

[3] Embrapa Cerrados, PO Box 08223, Brasília CEP 73310-970, Distrito Federal, Brazil

[4] Current address: Dupont Pioneer, Palmas, Tocantins, Brazil

[†] **Equal contributors.**

Abstract

Background

Brachiaria ruziziensis is one of the most important forage species planted in the tropics. The application of genomic tools to aid the selection of superior genotypes can provide support to

B. ruziziensis breeding programs. However, there is a complete lack of information about the *B. ruziziensis* genome. Also, the availability of genomic tools, such as molecular markers, to support *B. ruziziensis* breeding programs is rather limited. Recently, next-generation sequencing technologies have been applied to generate sequence data for the identification of microsatellite regions and primer design. In this study, we present a first validated set of SSR markers for *Brachiaria ruziziensis*, selected from a *de novo* partial genome assembly of single-end Illumina reads.

Results

A total of 85,567 perfect microsatellite loci were detected in contigs with a minimum 10X coverage. We selected a set of 500 microsatellite loci identified in contigs with minimum 100X coverage for primer design and synthesis, and tested a subset of 269 primer pairs, 198 of which were polymorphic on 11 representative *B. ruziziensis* accessions. Descriptive statistics for these primer pairs are presented, as well as estimates of marker transferability to other relevant *Brachiaria* species. Finally, a set of 11 multiplex panels containing the 30 most informative markers was validated and proposed for *B. ruziziensis* genetic analysis.

Conclusions

We show that the detection and development of microsatellite markers from genome assembled Illumina single-end DNA sequences is highly efficient. The developed markers are readily suitable for genetic analysis and marker assisted selection of *Brachiaria ruziziensis*. The use of this approach for microsatellite marker development is promising for species with limited genomic information, whose breeding programs would benefit from the use of genomic tools. To our knowledge, this is the first set of microsatellite markers developed for this important species.

Background

The area planted with forage crops in the tropics extends for hundreds of millions of hectares. In Brazil alone, the forage cropped land exceeds 100 M ha [1], where four *Brachiaria* species (*B. brizantha, B. decumbens, B. ruziziensis* and *B. humidicola*) cover 85% of the cultivated pastures [2]. Only a few apomictic *Brachiaria* clones occupy tens of millions of hectares in the country [3], what represents a high risk of genetic vulnerability for forage production. This risk could be reduced with the increased use of genetic diversity conserved in

germplasm banks in order to generate recombinant genotypes in breeding programs. The development and adoption of new *Brachiaria* cultivars with a broad genetic base is crucial for the diversification of forage pasture in the tropics. The development of new cultivars must be a dynamic process, providing the pasture production sector with increasing genetic diversity.

Among the four *Brachiaria* species most cultivated in Brazil, ruzigrass (*Brachiaria ruziziensis*, 2n=2x=18) stands out as a diploid species with sexual reproduction. Polyploid *Brachiaria* species such as *B. brizantha, B. decumbens* and *B. humidicola* typically present apomictic reproduction, a disadvantage for breeding programs that rely on sexual crosses and recombination for superior genotype selection. Ruzigrass has good forage quality, fast growth in the beginning of the rainy season and is readily adaptable to forest-crop-livestock integration systems, not only for animal feeding (green pasture or hay) but also as soil coverage for no-till farming. After tetraploidization, ruzigrass plants can be crossed with other *Brachiaria* species, making the inter-specific introgression of genes possible. Seed production is uniform, since flowering occurs only once a year. This favors a decrease in seed production costs and an increase in seed quality. The elimination of the seed shattering trait is an essential move in enabling full domestication of *B. ruziziensis*, and will contribute to production of high quality seeds, turning *B. ruziziensis* into an essentially agricultural crop.

Ruzigrass has a relatively small genome (~600 Mbp [4]), similar to other model cereal species, such as rice (430 Mbp) and sorghum (700 Mbp). This enables genome analysis initiatives and the development of molecular tools to support breeding programs. In contrast, tetraploid *Brachiaria* species (e.g. *B. decumbens, B. brizantha*) have larger and more complex genomes (> 1,600 Mbp). Therefore, ruzigrass has great potential to be used in breeding programs for pasture diversification, especially in combination with genomic tools aiding the selection of superior genotypes.

The employment of these genomic tools would favor a more dynamic development of new cultivars for this species. However, there is a lack of information about the *B. ruziziensis* genome. Little or nothing is known about the number of genes, distribution of gene families, abundance and diversity of retro-elements, QTL localization of traits of economic importance, genome collinearity with model species, or abundance of repetitive sequences. Genomic tools, such as molecular markers (e.g. microsatellites and SNPs), to support breeding programs are simply not available.

Traditional methods for the identification of microsatellite markers usually demand the construction of small-insert genomic libraries, colony selection by microsatellite-containing probe hybridization, sequencing of selected clones, primer design for suitable flanking regions, and assessments on the marker polymorphism by PCR analysis on a germplam sample. Later on, methods employing microsatellite-enriched genomic libraries diminished costs, time and workload necessary for marker development [5-7].

More recently, research groups have been applying next-generation sequencing technologies to generate sequence data for the genome identification of microsatellite regions and primer design [8-12]. For this purpose, both genomic DNA and genic regions (using cDNA libraries) have been used as templates for sequencing. The impact of this approach on microsatellite maker development is evident: partial genomic surveys using even fractions of a lane on next-generation sequencing machines allow the discovery of thousands of potentially amplifiable microsatellite regions which can be selected for primer design [13]. This is a promising approach for species with limited genomic information, whose breeding programs would greatly benefit from the use of genomic tools.

In *Brachiaria*, marker development initiatives so far used microsatellite enriched libraries to obtain SSRs for the species *B. brizantha* [14-16] and *B. humidicola* [14,17,18]. In summary, around 28 markers were polymorphic in *B. brizantha*, and 65 in *B. humidicola*. These authors tested the transferability of these markers to other *Brachiaria* species, and the rates of successful amplifications varied with the target species. At least 12 out of the 28 markers developed from *B. brizantha* produced amplified PCR products in *B. ruziziensis* DNA. Similarly, PCR products were observed on 13 out of 65 microsatellites developed from *B. humidicola*, when these were tested on *B. ruziziensis* DNA. No information on descriptive statistics such as polymorphic information content (PIC), allelic variation or heterozygosity estimates has been provided for these markers when tested on ruzigrass accessions.

In this study, we present a first set of 500 SSR markers developed for *Brachiaria ruziziensis*, selected from a *de novo* partial genome assembly of single-end Illumina reads. Descriptive statistics for 198 of these markers are provided. A set of 11 multiplex panels for the simultaneous amplification of the 30 most informative markers (ranked by their Polymorphism Information Content) is made available. These markers will be readily useful for the *B. ruziziensis* breeding program, aiding in areas such as germplasm characterization,

construction of linkage and QTL maps, gene flow and mating system evaluation, and marker assisted selection.

Results

Number of SSR loci initially detected in the ruzigrass genome

We restricted our search for microsatellite-containing regions to perfect di- tri- and tetranucleotide motifs only. After partial *de novo* genome assembly, a total of 139,098 perfect microsatellite loci were detected (Table 1). In order to select loci for subsequent primer design, we looked for perfect microsatellites in contigs >200 pb with a minimum 10X coverage. This reduced the number of regions to 85,567.

Table 1 Summary of Illumina single-end read sequence data and *de novo* assembly; perfect di-, tri- and tetra-nucleotide SSR loci for *Brachiaria ruziziensis*

	All contigs	**Only contigs >200 bp**
Reads #	186,764,108	186,764,108
Read average length bp	76	76
Reads bp	14,194,072,208	14,194,072,208
Mapping Parameters (LF - SIM)	0.5 - 0.8	1.0 - 1.0
Reads Matched	179,690,233	68,644,823
Matched bp	13,656,457,708	5,217,006,548
Contigs #	1,113,797	419,751
N50	585	954
Contigs bp	367,553,010	277,588,081
Average coverage	37x	18,8x
Contig average length	330	661
Perfect microsatellite sequences	139,098	85,567
Di-nucleotides	13,127	3,919
Tri-nucleotides	113,098	72,902
Tetra-nucleotides	12,892	8,746

Most frequent motif types and repeat numbers

Tri-nucleotide repeats were the most abundant class of microsatellites (72,902 regions) detected in the partially assembled ruzigrass genome, followed by tetra-nucleotide (8,746)

and di-nucleotide repeats (3,919) (Figure 1A). AG, CCG, and AAAT were the most frequent types of microsatellite sequences detected on each class (Figure 1B). The most frequent tri-nucleotide repeat motif (CCG) was particularly the most abundant one, comprising 19.8% of the perfect microsatellite regions detected on contigs with at least 10X coverage. Di- and tetra-nucleotide repeat motifs, on the other hand, had a more balanced distribution among different classes. The average number of repeats was three for tri- and tetra-nucleotides, and six for di-nucleotides.

Synthesized primer pairs

A total of 1,135 perfect microsatellite loci were detected in contigs with a minimum 100X coverage. We selected 500 loci at random for primer design and synthesis, which were given the "Brz" prefix. Additional file 1: Table S1 includes information regarding their forward and reverse primer sequences, their melting temperatures, repeat motifs, and expected product sizes. A subset of these loci was labeled with fluorescent dyes and multiplexed in order to test their efficiency on genotyping ruzigrass accessions. We tested 92 multiplex panels containing 269 primer pairs (panels contained up to three loci). Successful genotyping of 239 of these loci was achieved, while the remaining 30 loci presented either difficult interpretation of genotyping data, or absence of amplified products. However, no PCR optimization attempts were made for these loci. This represents a minimum 88.9% success rate of PCR amplification in unoptimized conditions for microsatellite loci generated from this partial *de novo* genome assembly. Among those 239 markers presenting coherent, interpretable amplified products, 198 (82.8%) markers were polymorphic when tested on 11 diverse African-derived ruzigrass accessions. If we consider the loss of microsatellite markers in the whole process, at least 73.60% of the 269 tested loci represent polymorphic, informative markers which can be readily applied to ruzigrass germplasm characterization and breeding. Figure 2 shows an example of electropherogram for one of the tested panels on three ruzigrass accessions.

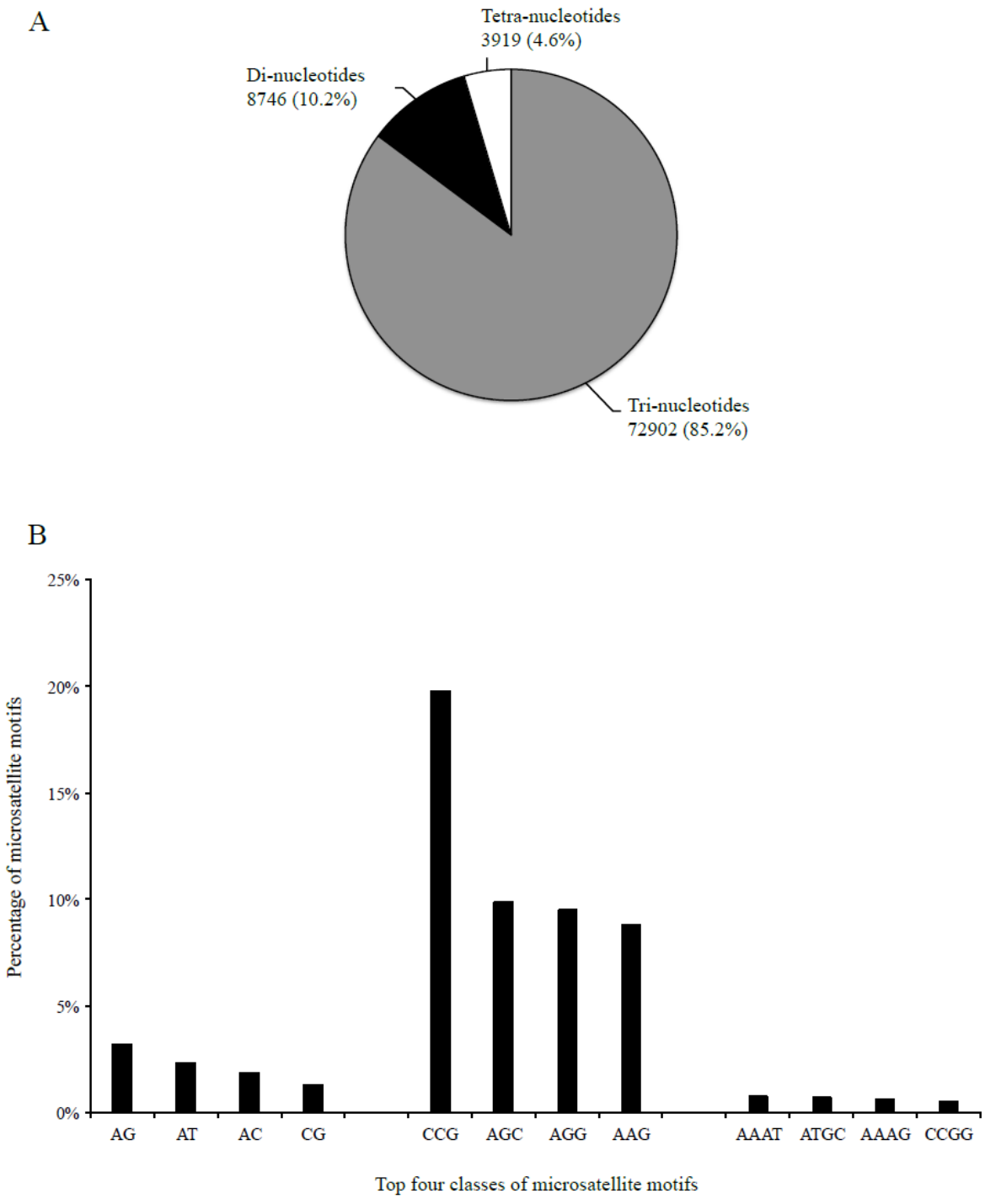


Figure 1 (a) Distribution of di-, tri-, and tetra-nucleotide microsatellites on contigs with a minimum 10X coverage; (b) Distribution of most frequent repeat motifs on contigs with a minimum 10X coverage.

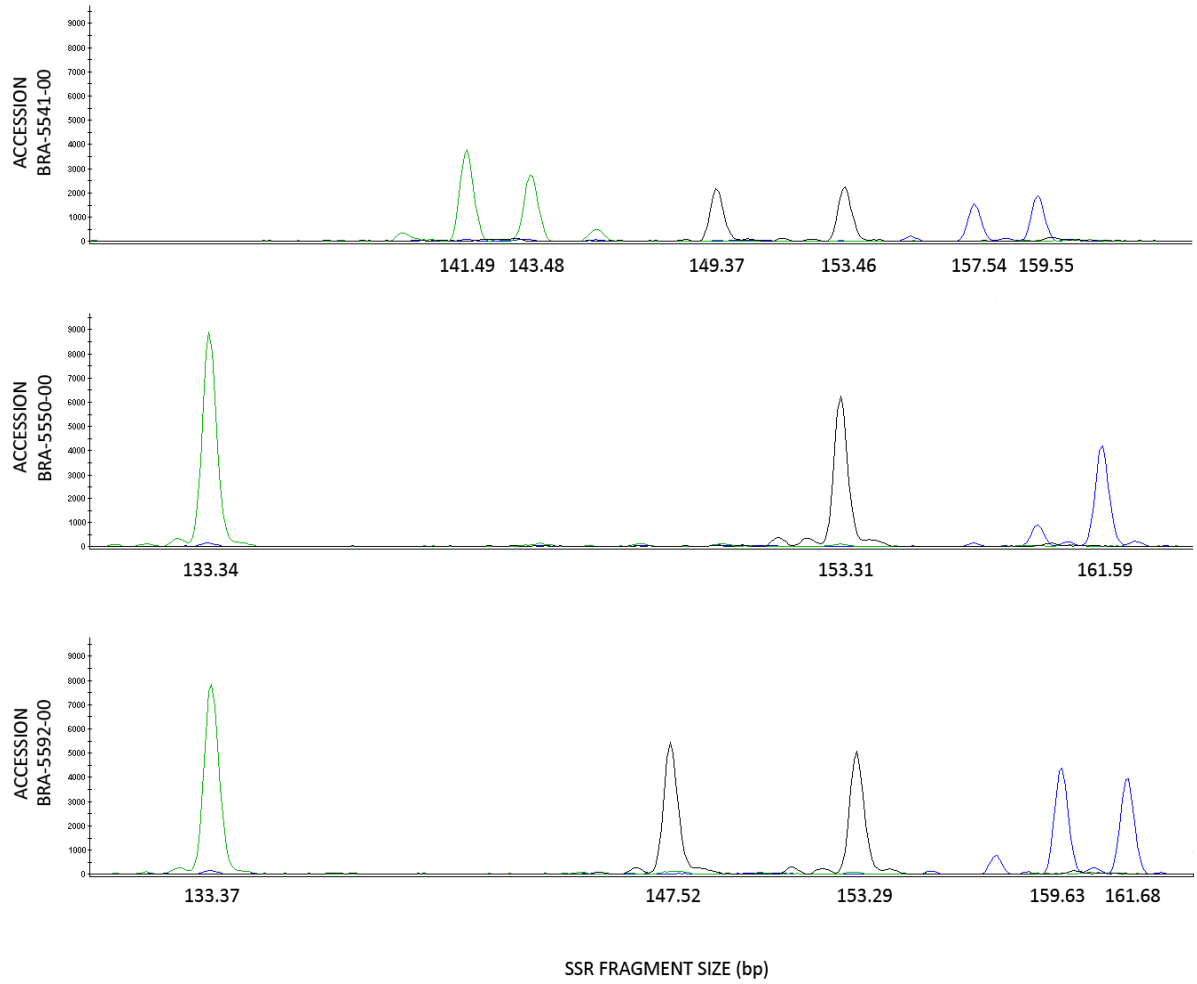


Figure 2 Electropherograms of a mutiplex panel showing amplification patterns of three Brz markers (Brz0059, green; Brz0069, black; Brz0047, blue), in three ruzigrass accessions (BRA-5541-00, BRA-5550-00, and BRA-5592-00).

Descriptive statistics for each SSR marker

Genotyping of 11 ruzigrass accessions with these 198 markers detected 835 alleles. The initial database of allele frequencies in *Brachiaria ruziziensis* shows 8.38% of rare alleles (with a frequency < 0.05), 64.07% of intermediate alleles (0.05 < frequency < 0.30) and 27.54% of abundant alleles (frequency > 0.30). Additional file 2: Table S2 presents the descriptive statistics information regarding these polymorphic markers.

The number of observed alleles for all polymorphic SSR markers ranged from 2 to 12, with an average value of 4.22 alleles per locus. Their expected heterozygosity (He) values ranged from 0.09 to 0.84, with an average of 0.518. Observed heterozygosity (Ho) values ranged

from 0 to 1, with an average of 0.410. The Polymorphism Information Content (PIC) values ranged from 0.09 to 0.87, with an average of 0.519.

Expected product sizes for each microsatellite marker are based on sequence information generated by the *de novo* assembly process. We checked whether the size ranges for the polymorphic loci included their expected product size. This was true for 70.2% of the loci (139 out of 198). The proportion of markers that generated amplicons within 10% of their expected sizes was 95.9% (190 out of 198). No markers presented amplicons 90% larger or smaller than expected.

We ranked the 30 most informative markers regarding their PIC values and assembled them into 11 multiplex panels for fast ruzigrass genotyping. The average PIC value for the 30 markers was 0.803, varying from 0.74 to 0.87. Table 2 shows these panels and markers and their respective primer sequences and descriptive statistics.

Table 2 A set of 11 multiplex panels including the 30 most informative ruzigrass microsatellite markers

Panel	Marker	Dye	Forward Primer	Reverse Primer	Allele No	Observed size ranges	He	Ho	PIC
1	Brz0182	NED	ACGTTATTGGACTTGGGTGA	AGCCTGACCAAATTCTTGTG	10	252-328	0.823	0.545	0.868
	Brz0097	HEX	TAATTTGTTCCACCCACAGG	GTGACAGAGTTCGGGAGCTA	5	234-242	0.705	0.375	0.747
2	Brz0075	6-FAM	GAAGCTGCAAAGGCTGAGT	GGAGGAGAGAGAAGAGCAAGA	8	129-153	0.809	0.727	0.839
	Brz0148	6-FAM	GCTCTTGACCTTGACGATGT	TGCACTTGAGAGAGACGAAA	8	248-274	0.787	0.909	0.800
	Brz0083	HEX	CATGATATTTGCCTGTCAAGG	AGCACCGGTGATGTGAATA	6	233-249	0.765	0.778	0.788
3	Brz0017	HEX	TTCCATTTATTTGCCTGTTCA	ATTTTCCCTATCCGACCTTTC	11	134-160	0.840	1000	0.864
	Brz0116	HEX	TCAAGAAATGGACTCCCAAA	TCTAGGTCATGCAAGCCATT	9	223-271	0.803	0.900	0.827
	Brz0047	6-FAM	TGTGAGACATAAACCATTGGAA	AATGGGTGCTGGAAATGTAAC	7	150-170	0.731	0.556	0.762
4	Brz0021	HEX	CAGCTGAAAGTTCCCAAAAAT	CTGAATGATAAAGGGTGCAAA	9	151-183	0.770	0.400	0.816
	Brz0087	NED	TTCCCCCACTACTCATCTCA	AACAGCACACCGTAGCAAGT	6	239-273	0.716	1000	0.748
5	Brz0065	6-FAM	AGCTAAGCAAATTTCAAGAACG	TAATGTGGAACATTGCCCTAA	12	130-166	0.829	0.700	0.875
	Brz0130	6-FAM	TCCTTTCATGAACCCCTGTA	CATCGCACGCTTATATGACA	9	242-266	0.820	0.636	0.858
	Brz0131	HEX	TGCAATGACATTAAATCAACC	GCTGCAACACAAACAAAATAA	6	254-264	0.712	0.714	0.744
6	Brz0147	HEX	CTGAGGACGCTCCTACTGAA	TTGATTTCAACACCCCAACT	10	240-288	0.825	0.700	0.868
	Brz0031	6-FAM	CCCCCATTTAACACCATAGTT	GCTCAAAATGCAATGTACGTG	7	144-156	0.770	0.667	0.804
7	Brz0177	6-FAM	TGGAGTTGAGGCTTTAGGAA	GTGTTTGGAAACCACTTGCT	6	291-319	0.725	0.125	0.795
	Brz0107	6-FAM	AGAGGAATTGACTTGGAAAAA	GCATGCACGTAAATTTTCACT	6	227-247	0.747	0.444	0.788
	Brz0004	6-FAM	TTGTTGTGGTACACCGGTACT	CAAAACCTGAATCACCATGTC	6	113-155	0.703	0.222	0.745
8	Brz0118	NED	AGGAGGTCCAAATCACCAAT	CGTCAGCAATTCGTACCAC	10	237-263	0.812	0.636	0.849
	Brz0219	HEX	GCAGTTCTTGCTTTTTCAGG	TCTCCTTATGCAAGGCTTC	6	294-304	0.768	0.818	0.778
	Brz0156	6-FAM	GCCATGATGTTTCATTGGTT	TTTTGCACCTTTCATTGCTT	7	239-265	0.752	0.636	0.770
9	Brz0142	6-FAM	GCTGGGTTATGCTAATGCAA	TCAAGCATGAACATTGAAACA	10	241-287	0.823	0.875	0.871
	Brz0180	HEX	CACACGGTCCATCTTGATTT	TCCATAATGCATTGTCTTGAAA	7	285-305	0.751	0.091	0.800
	Brz0089	NED	CAAACCTATTCCACGGTCAA	TGGACAATGCTATTCAAACG	7	224-248	0.710	0.571	0.759
10	Brz0048	HEX	GAATCTAAGCAGCGGATCAAT	TCACAAGAAGGTCCTCACAAG	9	139-161	0.813	0.818	0.839
	Brz0206	NED	GAAGTGGCAAGACACACACA	TGAGCTTTTCGTCTCTCCTG	7	278-302	0.757	0.600	0.783
	Brz0038	6-FAM	CTGAAAATAAGAGCCGTCCAT	ATAAGGTGAGCCACAACTGAG	6	140-154	0.772	0.909	0.778
11	Brz0171	6-FAM	TTGTCTCACTTGTGCACTCC	GCTAGCAGGTAGCAAGATGG	7	312-348	0.725	0.250	0.787
	Brz0015	6-FAM	AATAGAAAACGTGAGCCCATT	TCCACCAATATGATTCAAACG	6	144-156	0.764	0.636	0.783
	Brz0152	NED	ATGCTGCACTTACTGGTTCA	GGCTATCAATTCGAAGACCA	6	228-248	0.748	0.667	0.774

Transferability to other *Brachiaria* species

A survey on the potential transferability of microsatellite markers generated for ruzigrass to other *Brachiaria* species showed that 90.9% of the 198 polymorphic markers presented amplified PCR products on *Brachiaria brizantha* cv. Marandu, 67.7% on *B. brizantha* cv. Piatã, and 87.9% on *B. brizantha* cv. Xaraés. The percentage of potentially transferable markers to *B. decumbens* cv. Basilisk was 92.9%. Finally, for *Brachiaria humidicola* cv. Tupi, only 42.9% of 198 markers showed amplified PCR products.

Discussion

A true revolution is taking place on our ability to identify and develop microsatellite markers either for breeding, germplasm characterization, or conservation. The steady decrease in costs for obtaining next-generation sequencing data has made possible for research groups with access to an NGS facility to put a new model of microsatellite development to the test.

Most of the first published papers reporting the use of next-generation sequencing technologies for the development of microsatellite markers used either shotgun pyrosequencing of genomic DNA [8-12], or of enriched libraries [19]. Illumina sequencing was first applied to transcriptome sequencing and assembly, followed by the detection of genic SSR markers [20,21]. Castoe et al. [13] tested the use of Illumina paired-end reads of genomic DNA, without enrichment or assembly of reads, to detect potentially amplifiable microsatellite loci on three different organisms. This approach was also used by O'Bryhim et al. [22] to develop microsatellite markers on an endangered scaleshell species. Castoe's work does not present any data on the test of synthesized primer pairs. O'Bryhim's paper reports the test of 48 primer pairs, 16 of which were polymorphic.

We show that reads from an Illumina single-end run, when assembled *de novo* with high levels of stringency, are also suitable for the identification of microsatellite regions. Even though we haven't tested Castoe's scripts to detect potentially amplifiable loci from unassembled reads, we believe the assembly process adds a consistent level of sequence quality. That increases the chance of finding good-quality flanking regions for which primer pairs can be designed.

Squirrel et al. [23] used the term “attrition rate” to describe the loss of loci at each step of microsatellite marker development. For traditional projects - which include the construction of clone libraries, the sequencing of clones, microsatellite identification, primer design, and PCR - their estimate based on a review of published papers showed that, on average, 83% of the sequenced clones would be lost due to problems in different steps of the development process.

The application of this criterion to measure how much effort is necessary to develop functional, polymorphic microsatellite markers using genome surveys based on next-generation sequencing depends on the definition of what initial count is used. In our case, depending on the imposed stringency on contig coverage, our initial number of potentially useful, perfect microsatellite markers ranged from 139,098 to 85,567 (at least 10X contig coverage), and finally to 1,135 (100X coverage). If we chose the most stringent parameter, we would expect that from our 1,135 microsatellite-containing sequences, 729 would be suitable for primer synthesis (46% of mean attrition rate on this step), and 365 would be polymorphic (50% mean attrition rate). If we only consider that final step, the expected number of functional polymorphic markers from our set of tested primer pairs would be 135 (starting with our 269 loci). Our observed number of polymorphic markers was higher, 198 of our 269 tested primer pairs were polymorphic (73.6%).

We could apply the attrition rate estimates published by Squirrel et al. to answer one more question: given our final set of functional polymorphic microsatellite loci, how much effort would be necessary in previous steps of marker development if we were using a traditional clone library approach? The answer is that in order to obtain 198 functional polymorphic loci, 1,146 clones from an enriched library would have to be sequenced, 733 microsatellites would have to be identified, and 396 primer pairs would have to be synthesized and tested.

It is obvious that when comparing these estimates, factors such as the abundance of microsatellite regions on the genome of interest are taken for granted. For practical purposes, a more useful comparison would be that between a clone library sequencing method and a next-generation sequencing method on the same organism. In this case, not only the final number of useful markers would be considered, but also costs, time and laboratory workload. Santana et al. [19] have done that for the fungus *Fusarium circinatum*, a pine pathogen. While a single Roche 454 run using pooled ISSR-PCR products detected 231 potentially amplifiable microsatellites

(out of 1,692 contigs and singletons), Sanger sequencing of 100 clones containing ISSR-PCR fragments allowed the detection of 8 potentially amplifiable sequences.

We can compare our effort with previous microsatellite development initiatives for other *Brachiaria* species. In *B. brizantha* [15], 96 clones from an enriched library were sequenced, 19 primer pairs were designed and tested, and 13 of those were polymorphic. A new set of 15 polymorphic primers for this species was published by Vigna et al. [16], using the same enriched library. For *B. humidicola*, 384 clones were sequenced, 38 primer pairs were tested, and 27 were polymorphic [17]. A new set of 40 primer pairs was tested by Vigna et al. [18], 38 of which were polymorphic. No microsatellite markers had been developed so far for *Brachiaria ruziziensis*.

It seems, therefore, that the detection and development of microsatellite markers from genome assembled Illumina single-end DNA sequences is highly efficient. This approach should be especially considered for species with limited genomic information.

The need for further germplasm collection expeditions to increase the genetic diversity of *B. ruziziensis* kept in germplasm banks should also be mentioned. It was observed that roughly 30% of the expected allele sizes were not detected on the 11 ruzigrass accessions genotyped in this study. Since the plant used to generate the single-end sequences is derived from a self-pollinated plant collected in the field in Brazil, this data indicates that there is genetic variation in ruzigrass that is out of the allele variation boundaries observed in the analyis of the 11 African-derived genotypes used in this experiment. It is possible that new germplasm collection initiatives in pastures established in the 1960-1970's in Brazil will identify accessions with useful genetic diversity for ruzigrass breeding programs.

Finally, although we consider the data on transferability of ruzigrass microsatellite markers to other *Brachiaria* species rather preliminary, the higher proportion of successful PCR amplifications on *B. brizantha* and *B. decumbens* cultivars indicates a closer phylogenetic distance between these species and *B. ruziziensis*, when compared with *B. humidicola*.

Conclusions

We show that the detection and development of microsatellite markers from genome assembled Illumina single-end DNA sequences is highly efficient. The developed markers are readily suitable for genetic analysis and marker assisted selection of *Brachiaria ruziziensis*. The use of this approach for microsatellite marker development is promising for species with limited genomic information, whose breeding programs would benefit from the use of genomic tools. To our knowledge, this is the first set of microsatellite markers developed for this important species.

Methods

Sequencing and *de novo* partial assembly of the *B. ruziziensis* genome

B. ruziziensis genome sequencing was performed with DNA extracted from a self-pollinated plant (FSS-1 clone), in order to increase homozygosity and, as a consequence, facilitate the *de novo* genome assembly. Sequencing was performed from a genomic DNA fragment library, amplified by cluster generation by bridge PCR, allowing the massive parallel sequencing by synthesis in an Illumina GAII sequencer. Assembly routines were performed on CLC Genomics Workbench software (CLC Bio, Aarhus, Denmark). An assembly mapping was obtained after removing of Illumina adapters and low quality sequences using the CLC trimmer function (default limit = 0.05). The assembly procedure used the parameters Length Fraction (LF) and Sequence Similarity (SIM) between DNA reads, as described by the CLC Genomics Workbench software, with maximum stringency (0.50 LF and 0.80 SIM). The minimum contig length parameter was set to 70 bp.

Selection criteria for microsatellite loci in *B. ruziziensis*

Microsatellite sequence discovery was carried with Phobos [24]. Initially, we searched for di-, tri-, and tetra-nucleotide loci with perfect repeat motifs on assembled contigs with at least 10X coverage. This allowed a preliminary survey of the most frequent types of repeat motifs on the assembled genome, and the number of repeat motifs for the detected loci. A dataset with contigs >200 bp was then used to map the reads using maximum stringency (100% LF and 100% SIM), in order to minimize the error of consensus sequences while improving the coverage of

conserved sequences. With this procedure, the average length of resulting contigs was increased. Perfect microsatellites which occurred in the contigs greater than 200 bp and with coverage above 10x could be recovered using Phobos. A final set of 500 microsatellites with minimum 100x coverage was then selected for analysis and validation (Additional file 1: Table S1). The microsatellite containing sequences received the GeneBank accession numbers KC181352 - KC181851.

In order to test some of these loci on *Brachiaria ruziziensis* germplasm, primer pairs were designed with Primer3Plus [25]. From the initial list of detected microsatellites, we generated a subset of loci which were present on contigs with at least 100X coverage. Two hundred and seventy primer pairs were designed (240 di-nucleotides, 20 tri-nucleotides, and 10 tetra-nucleotides). Fluorescent labels were added to the forward oligos of each primer pair so that multiplexing and genotyping would be performed on an automated DNA sequencer.

Plant material for SSR genotyping

We tested the synthesized primer pairs on eleven ruzigrass samples - ten accessions from the Embrapa Germplasm Collection and one cultivar (Kennedy). The ruzigrass accessions were selected for this study based on their expected high genetic diversity, since they are progenies of original germplasm accessions collected in the 1980's in different countries of Africa, where *B. ruziziensis* is endemic [26]. Seeds were germinated and DNA was extracted using a standard CTAB protocol [27] with modifications, as described in [28]. Leaves from five cultivars of other *Brachiaria* species were also collected and had their DNA extracted. These were cultivars Marandu, Piatã and Xaraés (*Brachiaria brizantha*), cultivar Basilisk (*Brachiaria decumbens*) and cultivar Tupi (*Brachiaria humidicola*), all of them registered for commercial cultivation in Brazil. They were genotyped in order to test the transferability of SSR markers designed for *B. ruziziensis* to commercially important polyploid *Brachiaria* species. DNA concentrations were measured on a Nanodrop 2000 spectrophotometer (Thermo Scientific, USA), and samples were diluted on TE buffer pH 8.0 to a concentration of 2 ng/μL.

Genotyping using multiplex panels of SSR markers

Multiplex panels were designed using Multiplex Manager [29]. They included up to three loci per panel, and all loci in each panel had the same microsatellite repeat motif size. PCR's were carried in a final volume of 5 μL containing 2 ng of genomic DNA, 1X QIAGEN Multiplex PCR Kit Master Mix (QIAGEN), 0.5X Q-Solution (QIAGEN), and 0.2 μM of each primer. Reactions were performed on a Veriti™ Thermal Cycler (Applied Biosystems, USA) using the following amplification program: 95°C for 15 minutes; 30 cycles at 94°C for 30 seconds, 52°C for 90 seconds, and 72°C for 60 seconds; a final extension step at 60°C for 60 minutes. PCR products were diluted with an equal volume of Milli-Q water, added 10 μL of Hi-Di™ Formamide (Applied Biosystems, USA), a ROX-labeled internal size standard, and denatured at 94°C for 5 minutes. Denatured products were injected on an ABI 3730 (Applied Biosystems, USA) automated sequencer. Allele size calling and genotyping were carried with the GeneMapper® Software v4.1 (Applied Biosystems, USA). Automated allelic binning was performed with AlleloBin [http://www.icrisat.org/bt-software-d-allelobin.htm], which is based on an algorithm described in [30]. PowerMarker v. 3.25 [31] was used to generate a table of summary statistics for all loci, as well as a database of allelic frequencies.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PITS and AMM prepared genomic libraries, worked on genome assembly, detection of microsatellite sequences, primer design, and multiplex panel development. EGG genotyped ruzigrass accessions with microsatellite markers, analyzed genotyping data and performed statistical analyses. MPF helped analyze genotyping data, performed statistical analyses, selected loci for multiplex panels and drafted the manuscript. MEF conceived of and surpervised the study, performed statistical analyses and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Fausto Souza Sobrinho, Claudio Takao Karia and Marcelo Ayres Carvalho for providing the *Brachiaria* accessions used in this work. This research was sponsored by EMBRAPA Macroprograma 2 – Grant # 02.12.02.002.00.00.

References

1. de Lima M, Pessoa M, Neves M, de Carvalho E: **Emissões de metano por fermentação entérica e manejo de dejetos de animais**. In *Segundo inventário brasileiro de emissões e remoções antrópicas de gases de efeito estufa*. Brasília: Ministério da Ciência e Tecnologia; 2012:120.
2. Barcellos AO, Vilela L, Lupinacci AV: **Produção animal e pasto: desafios e oportunidades**. In *Encontro Nacional do Boi Verde - A Pecuária Sustentável: 2001*. Uberlândia: Sindicato Rural de Uberlândia; 2001:29–64.
3. Barbosa RA: *Morte de pastos de braquiárias*. Campo Grande: Embrapa Gado de Corte; 2006.
4. Ishigaki G, Gondo T, Ebina M, Suenaga K, Akashi R: **Estimation of genome size in *Brachiaria* species.** *Grassl Sci* 2010, **56**(4):240–242.
5. Billotte N, Lagoda PJL, Risterucci AM, Baurens FC: **Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops.** *Fruits* 1999, **54**(4):277–288.
6. Ostrander EA, Jong PM, Rine J, Duyk G: **Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences.** *Proceedings Of The National Academy Of Sciences Of The United States Of America* 1992, **89**(8):3419–3423.
7. Paetkau D: **Microsatellites obtained using strand extension: an enrichment protocol.** *Biotechniques* 1999, **26**(4):690–692. 694–697.
8. Abdelkrim J, Robertson B, Stanton J-A, Gemmell N: **Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing.** *Biotechniques* 2009, **46**(3):185–192.
9. Castoe TA, Poole AW, Gu W, de Koning AP J, Daza JM, Smith EN, Pollock DD: **Rapid identification of thousands of copperhead snake (Agkistrodon contortrix) microsatellite loci from modest amounts of 454 shotgun genome sequence.** *Mol Ecol*

Resour 2010, **10**(2):341–347.

10. Csencsics D, Brodbeck S, Holderegger R: **Cost-effective, species-specific microsatellite development for the endangered Dwarf Bulrush (Typha minima) using next-generation sequencing technology.** *J Hered* 2010, **101**(6):789–793.
11. Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, Seehalak W, Sommanas W, Tragoonrung S, Srinives P: **Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (Vigna radiata (L.) Wilczek).** *BMC Plant Biol* 2009, **9**:137.
12. Zhu H, Senalik D, McCown BH, Zeldin EL, Speers J, Hyman J, Bassil N, Hummer K, Simon PW, Zalapa JE: **Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (Vaccinium macrocarpon Ait.).** *Theor Appl Genet* 2012, Jan, 124(1):87-96.
13. Castoe TA, Poole AW, de Koning APJ, Jones KL, Tomback DF, Oyler-McCance SJ, Fike JA, Lance SL, Streicher JW, Smith EN, *et al*: **Rapid microsatellite identification from illumina paired-end genomic sequencing in two birds and a snake.** *PLoS One* 2012, **7**(2):e30953.
14. Cançado LJ: *Caracterização da diversidade genética molecular em germoplasma de Brachiaria spp*. Campinas: Universidade Estadual de Campinas; 2009.
15. Jungmann L, Sousa ACB, Paiva J, Francisco PM, Vigna BBZ, do Valle CB, Zucchi MI, DE Souza AP: **Isolation and characterization of microsatellite markers for *Brachiaria* brizantha (Hochst. ex A. Rich.) Stap.** *Conserv Genet* 2009, **10**(6):1873–1876.
16. Vigna BBZ, Jungmann L, Francisco PM, Zucchi MI, Valle CB, Souza AP: **Genetic diversity and population structure of the *Brachiaria* brizantha germplasm.** *Tropical Plant Biology* 2011, **4**(3–4):157–169.
17. Jungmann L, Vigna BBZ, Paiva J, Sousa ACB, do Valle CB, Laborda PR, Zucchi MI, DE Souza AP: **Development of microsatellite markers for *Brachiaria* humidicola (Rendle) Schweick.** *Conserv Genet Resour* 2009, **1**(1):475–479.
18. Vigna BB, Alleoni GC, Jungmann L, do Valle CB, de Souza AP: **New microsatellite markers developed from *Urochloa* humidicola (Poaceae) and cross amplification in different *Urochloa* species.** *BMC research notes* 2011, **4**:523.
19. Santana Q, Coetzee M, Steenkamp E, Mlonyeni O, Hammond G, Wingfield M, Wingfield B:

Microsatellite discovery by deep sequencing of enriched genomic libraries. *Biotechniques* 2009, **46**(3):217–223.

20. Garg R, Patel RK, Tyagi AK, Jain M: **De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification.** *DNA research: an international journal for rapid publication of reports on genes and genomes* 2011, **18**(1):53–63.
21. Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X: **Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers.** *BMC Genomics* 2011, **12**:451.
22. O'Bryhim J, Chong JP, Lance SL, Jones KL, Roe KJ: **Development and characterization of sixteen microsatellite markers for the federally endangered species: Leptodea leptodon (Bivalvia: Unionidae) using paired-end Illumina shotgun sequencing**. *Conservation Genetics Resources*; 2012; 4(3):787-789.
23. Squirrell J, Hollingsworth P, Woodhead M, Russell J, Lowe A, Gibby M, Powell W: **How much effort is required to isolate nuclear microsatellites from plants?** *Mol Ecol* 2003, **12**(6):1339–1348.
24. Mayer C: *Phobos*. 3.3.11 edn; 2006–2010.
25. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM: **Primer3Plus, An enhanced web interface to Primer3.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W71–W74.
26. Keller-Grein G, Maas BL: **Natural variation in *Brachiaria* and existing germplasm collections**. In *Brachiaria: biology, agronomy and improvement*. Edited by Miles J, Maas BL, Valle CB. Cali: CIAT; 1996:16–42.
27. Doyle JJ, Doyle JL: **A rapid DNA isolation procedure for small quantities of fresh leaf tissue.** *Phytochemical Bulletin* 1987, **19**(1):11–15.
28. Ferreira ME, Grattapaglia D: *Introdução ao uso de marcadores moleculares em análise genética*. Brasilia: Embrapa-SPI; 1998.
29. Holleley CE, Geerts PG: **Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR.** *Biotechniques* 2009, **46**(7):511–517.
30. Idury RM, Cardon LR: **A simple method for automated allele binning in microsatellite markers.** *Genome Res* 1997, **7**(11):1104–1109.

31. Liu K, Muse SV: **PowerMarker: an integrated analysis environment for genetic marker analysis.** *Bioinformatics* 2005, **9**(21):2128–2129.

Additional files

Additional_file_1 as XLS

Additional file 1 List of 500 Brz markers, including their primer sequences, melting temperatures, expected product sizes, and repeat motifs

Additional_file_2 as XLS

Additional file 2 Descriptive statistics for 198 polymorphic ruzigrass markers, and information on their transferability to other *Brachiaria* species

Additional_file_1

List of 500 Brz markers, including their primer sequences, melting temperatures, expected product sizes, and repeat motifs

Additional_file_1

List of 500 Brz markers, including their primer sequences, melting temperatures, expected product sizes, and repeat motifs

SSR Code	Orientation	tm	Seq	Expected Product size	Motif
Brz0001	FORWARD	56.96	GCTGAACTAAACATTGGAGGA	154	(TC)8
	REVERSE	57.93	AAGTGTGTGCTTTTCACTTGG		
Brz0002	FORWARD	58.01	AGGATTGCAACAATGTGGTTA	152	(GA)8
	REVERSE	57.02	TCTTGGTAGGAGAGATGGTCTT		
Brz0003	FORWARD	58.81	AATCGGCACATCAAGAGAAGT	146	(AG)9
	REVERSE	60.59	CGCAAGAGCTCGACAGCTA		
Brz0004	FORWARD	57.93	TTGTTGTGGTACACCGGTACT	150	(AT)13
	REVERSE	57.91	CAAAACCTGAATCACCATGTC		
Brz0005	FORWARD	57.99	TGCAGGAGAAACACAAACTTC	149	(AT)5
	REVERSE	58.48	TTTGGCCATTTGGTTCTTAAT		
Brz0006	FORWARD	55.45	TGATACTTCTTATAACCGACAGC	145	(AT)5
	REVERSE	58.38	CAGCTAGCAAACGTCTCAAAA		
Brz0007	FORWARD	58.17	CAAGATTTTGAGGGGAGGTAA	158	(AC)9
	REVERSE	58.19	TCAACTCTGGCCTTTCTCTTT		
Brz0008	FORWARD	58.16	CGTGAACTTCTGCTGTGACTT	155	(CT)8
	REVERSE	58.33	GTAACGCTAAGCATGATGGTG		
Brz0009	FORWARD	53.96	GTAAGTACCATGTAAAAATGCAA	146	(TA)8
	REVERSE	58.02	AAGACAATAAGAGGCATGAGTGA		
Brz0010	FORWARD	57.91	GATTGAAATTGCTTGCCTGTA	160	(TA)7
	REVERSE	57.95	CAGTGAACACACCATAATCAATG		
Brz0011	FORWARD	57.71	AGCATAAGCACACAAATAAGCA	141	(CA)9
	REVERSE	58.30	CCTTAAGGTCCAGTCCTTTGA		
Brz0012	FORWARD	57.72	ACTCAAACAATCTCCAACACG	160	(AT)8
	REVERSE	57.71	CCCACAAATGGTGAATGTAAC		
Brz0013	FORWARD	58.06	TGATACTCACACAAGGGGAAA	155	(TA)9
	REVERSE	57.48	AAAAGACCCAATGAGAAAAGC		
Brz0014	FORWARD	57.97	CGCCACGGTTTCTTAGTCT	154	(CA)7
	REVERSE	58.24	TAAGGTACGGTGTGGCTAACA		
Brz0015	FORWARD	57.76	AATAGAAAACGTGAGCCCATT	155	(TA)10
	REVERSE	58.35	TCCACCAATATGATTCAAACG		
Brz0016	FORWARD	58.49	TATTGTGGAGGTGCATTTGTC	160	(AG)9
	REVERSE	57.82	CTGTCGTCGTAGAGATGGTGT		
Brz0017	FORWARD	58.15	TTCCATTTATTTGCCTGTTCA	146	(CT)12
	REVERSE	58.46	ATTTTCCCTATCCGACCTTTC		
Brz0018	FORWARD	58.16	TGCTGTGAATATTTCATTTCCA	157	(TA)6

	REVERSE	57.79	CAGGGACAACTAACACAGAACA		
Brz0019	FORWARD	57.97	GTCCTTTTCAAACACCCGTAT	152	(AT)6
	REVERSE	54.92	AGAGAAAATAAAAGCAAAGCAC		
Brz0020	FORWARD	58.17	GGAAAGAGATTCGGGTTGTTA	151	(CT)5
	REVERSE	58.19	CTACTCTCCCAGCCAGCTATC		
Brz0021	FORWARD	58.37	CAGCTGAAAGTTCCCAAAAAT	149	(CT)12
	REVERSE	57.77	CTGAATGATAAAGGGTGCAAA		
Brz0022	FORWARD	58.52	GTGTCATGCCATGTATGCTTT	154	(TA)5
	REVERSE	57.50	CTGGATCCATTAAACCACGTA		
Brz0023	FORWARD	57.80	ACTGAATTGCTTCCATCCTTT	144	(CA)15
	REVERSE	58.58	GGTACCCATGATGGTGAAGAT		
Brz0024	FORWARD	58.36	ATGTCTGGTGAGGGTTTGATT	156	(TA)8
	REVERSE	58.16	CTGGGAAAGATCAAAAGTGGT		
Brz0025	FORWARD	57.59	CACCTTTACACCTTGATTCCA	141	(CA)8
	REVERSE	58.25	CGACTTCGGTTGAAAACCTAT		
Brz0026	FORWARD	57.70	GCACCTTGTAACAATGCAAAT	160	(TA)5
	REVERSE	59.08	TCTTTGTGGATTTGGGTTAGC		
Brz0027	FORWARD	58.31	ACACGACGCAAATTCATTCTA	148	(AG)8
	REVERSE	57.94	CCTACAACGGTTATCCTCCAT		
Brz0028	FORWARD	57.72	CATGGACAAGGAGAAGATTGA	158	(TA)8
	REVERSE	54.84	TGGGAGTTAACATTAGTGTTTTT		
Brz0029	FORWARD	57.82	TTTGTGCCAAAGTCCAAATAG	150	(AG)14
	REVERSE	56.92	TATTCCAGCTTCTTCTGCCTA		
Brz0030	FORWARD	57.78	CCTTCCATGTTCACAGAAGAA	151	(CA)12
	REVERSE	57.99	TCACTTTGTTTCTTGCCTCAC		
Brz0031	FORWARD	57.74	CCCCCATTTAACACCATAGTT	157	(AT)9
	REVERSE	59.24	GCTCAAAATGCAATGTACGTG		
Brz0032	FORWARD	57.71	TCCTAGCAAAACGAGATCAGA	155	(AG)9
	REVERSE	57.73	CAACAATAGAGCGTTTGAAGC		
Brz0033	FORWARD	57.53	CCTTCATGGGTGAATCTGTAA	147	(CT)8
	REVERSE	57.76	TCTGTCACCAGGTTCTGTTTC		
Brz0034	FORWARD	57.77	CGGTGTTAATCATTCTGCACT	160	(AT)6
	REVERSE	56.31	TTGACCAACAGATTTGTTACCT		
Brz0035	FORWARD	58.43	GCCACTAATGAAAATCCCAAC	155	(CA)7
	REVERSE	57.77	CGTGGATGACACTTGCTTATT		
Brz0036	FORWARD	58.16	CAAGCCATTGATGAGATTGTC	144	(TC)9
	REVERSE	57.32	TCACCAAACTAGTGAGGGAAA		
Brz0037	FORWARD	57.30	GAAACTGCACAAAACACACAA	150	(AT)6
	REVERSE	54.00	AGGAACAATTTGAACCTAACA		
Brz0038	FORWARD	58.34	CTGAAAATAAGAGCCGTCCAT	152	(AG)9
	REVERSE	56.94	ATAAGGTGAGCCACAACTGAG		

Brz0039	FORWARD	58.04	ACTCGACTCCTTATGCGAGAT	150	(TA)6
	REVERSE	57.96	TTAACAGGTCTCATCGTCTGC		
Brz0040	FORWARD	58.56	ACCTCTTGTCCCTTGGTTACA	149	(TA)5
	REVERSE	58.49	GGAGATCGTTCAATTTGTTCC		
Brz0041	FORWARD	57.59	TGGACCTATGGCTGAATTATG	144	(AT)5
	REVERSE	57.80	ACTTGCTCAAGCGATAAGTGA		
Brz0042	FORWARD	58.13	CTTTTTATTGGAAGCCACCAT	152	(CA)9
	REVERSE	58.42	GGGTAAGGTAACCCCTATGCT		
Brz0043	FORWARD	57.75	TCATTCAGTCCTGGTGATAGC	149	(TA)8
	REVERSE	57.56	CATCAATCAATAGGTGCCACT		
Brz0044	FORWARD	58.31	TTCCTTTCTTTGCTTTGCTTT	272	(AG)9
	REVERSE	58.97	GCAACATTGCTGCAAATAGAA		
Brz0045	FORWARD	57.48	TTTCTTGGATCTAATTTTCATGC	144	(TC)7
	REVERSE	58.17	ACAGCAACCCACACGTATCTA		
Brz0046	FORWARD	59.21	TAAGCATTTCACTTCCCCTTG	160	(AT)6
	REVERSE	58.03	GGGTATAAGCCCATACAGACAA		
Brz0047	FORWARD	58.00	TGTGAGACATAAACCATTGGAA	153	(GA)8
	REVERSE	58.83	AATGGGTGCTGGAAATGTAAC		
Brz0048	FORWARD	58.43	GAATCTAAGCAGCGGATCAAT	149	(AG)13
	REVERSE	57.92	TCACAAGAAGGTCCTCACAAG		
Brz0049	FORWARD	57.56	GTCGGCCTTTCTAGATTCACT	154	(AT)6
	REVERSE	58.63	GGTTCTTTCACTGGACTCACC		
Brz0050	FORWARD	57.83	GCTATCCTAACTGGGGTGAAG	157	(AG)7
	REVERSE	58.80	AGACCCAGAAGGGAAGAGTTC		
Brz0051	FORWARD	51.02	CATAATTCTTAACTTGCTTAGTG	147	(TA)7
	REVERSE	57.57	AGATGAACTTCCCATCAAGGT		
Brz0052	FORWARD	57.84	TTGAGACAAAGTTCGTTGACC	159	(AT)5
	REVERSE	54.33	TCAGGTGTGAGTTAGTTTAGTGA		
Brz0053	FORWARD	58.04	GAGATCGCTGGAGACGAGT	150	(AG)9
	REVERSE	58.35	GATCCAAGATTTGTGGTTTCC		
Brz0054	FORWARD	57.52	CACATTGCAGATAGTGAAGCA	142	(AT)5
	REVERSE	58.37	TGGAAGGTGCTTGTAAGATGA		
Brz0055	FORWARD	57.05	AAGGTTAAAGCCCCTAAACAA	159	(AC)9
	REVERSE	57.84	TCCCAGCTTTCAATGTAGATG		
Brz0056	FORWARD	58.36	AACCCCAGTGTTTGTTATCGT	148	(AT)8
	REVERSE	57.38	ATTTATCACAAGCAACGAGGA		
Brz0057	FORWARD	57.77	ACAAGCTTTGCTCAGAAATGA	154	(GA)7
	REVERSE	57.79	AGTAGAAAGGCCTGCAGGTAG		
Brz0058	FORWARD	58.37	CGATCTGACAATGAAAACTGC	140	(CA)13
	REVERSE	57.26	TATACCGATTCACTGCACCTT		
Brz0059	FORWARD	57.74	GGAAAGAGGATAGCAATGACC	141	(TA)7

	REVERSE	58.28	TTAAAGTCCAATGCTTGTCCA		
Brz0060	FORWARD	57.84	AGATGAGGAAGACGAACAGGT	153	(AT)7
	REVERSE	57.21	GCTCAATCTCTCCTTCCTTTC		
Brz0061	FORWARD	58.17	TTGCCTGACAAGAAGTACAGC	151	(AT)6
	REVERSE	57.72	TTTGAGTGATCGTGTTTCACA		
Brz0062	FORWARD	57.50	TTGCGGTCAGCTTATAACAAT	155	(TC)6
	REVERSE	59.79	ATTGGGGAAAGATTTGAGCAT		
Brz0063	FORWARD	58.69	CAAACACTTGCAACCCAGATA	147	(CT)10
	REVERSE	57.36	CATTTTGGCTTTGATAATTGC		
Brz0064	FORWARD	58.09	TATGCAACTGTGTGCTGCTT	152	(CA)9
	REVERSE	57.67	TTGAAATAAATTCAGCCTCTTTG		
Brz0065	FORWARD	57.92	AGCTAAGCAAATTTCAAGAACG	147	(CA)11
	REVERSE	58.03	TAATGTGGAACATTGCCCTAA		
Brz0066	FORWARD	57.78	GTGACTGTGAGCAGGAACAAT	145	(TC)7
	REVERSE	58.06	CGAAAAATGAGAAGAGGAAGG		
Brz0067	FORWARD	56.74	TTAGATTCCTCAGGACATTGG	156	(AT)9
	REVERSE	58.28	TCCTATATGCCGTCGTACTCA		
Brz0068	FORWARD	57.07	TTGGTAGCTGTTGTTCCTCTC	152	(CT)18
	REVERSE	56.99	TCTGCAGACAATTGACAAAAA		
Brz0069	FORWARD	58.49	TGGAAGCAAGTTTCAGAATCA	160	(AC)9
	REVERSE	56.43	AAGTTAAAAAGACCTCGAAGGA		
Brz0070	FORWARD	58.05	CATGTCTGCTAGGCAGTGTTT	145	(AT)8
	REVERSE	58.09	AGAAGGTGACTTCCATTGACC		
Brz0071	FORWARD	57.92	ATTGCAGAAGTACATGCAAGG	142	(GA)8
	REVERSE	57.91	ATCACAACGACCACAACAGAT		
Brz0072	FORWARD	57.16	TGTATGCTTTATAGTGCCACAAG	140	(AT)8
	REVERSE	56.93	ATGGAGGCACTCTATTTCCTT		
Brz0073	FORWARD	57.97	CGAAAATCTAGCCAAACACAA	144	(TA)9
	REVERSE	58.08	CCAAAAGCCAAAATCTAAAGC		
Brz0074	FORWARD	58.51	GACGGGAGACCACTAATTCAC	147	(AC)5
	REVERSE	57.96	CATGGTAATTCCAATGTCTGC		
Brz0075	FORWARD	58.33	GAAGCTGCAAAGGCTGAGT	150	(CT)10
	REVERSE	57.94	GGAGGAGAGAGAAGAGCAAGA		
Brz0076	FORWARD	57.14	CCTAGAATGCGGAAGTAGTGA	151	(AT)7
	REVERSE	57.87	TTACGTGTTCCTCGACTCAAC		
Brz0077	FORWARD	58.13	ACCTCCTATCTTTCCATCGTG	141	(TC)9
	REVERSE	57.53	AACGAGCTCTATTAGAAGCATGA		
Brz0078	FORWARD	57.89	ACAATTCAAGAAGATGCGTTG	158	(AG)7
	REVERSE	57.29	GGAGTTCCCTGAGAGACAAAT		
Brz0079	FORWARD	57.83	AGAAGATCTTGCCTGAAAAGC	154	(TC)5
	REVERSE	58.05	TTCCTCATGGTATGGCATCTA		

Brz0080	FORWARD	57.85	ATTAAACTTGTGCAAGCATGG	150	(AT)5
	REVERSE	57.95	AAACAAGCATTTGCCCTTAGT		
Brz0081	FORWARD	58.15	TGTGAAGGGATTTCTTGCAT	247	(TC)5
	REVERSE	57.60	TTGTTTGCTGCTTATGTTGC		
Brz0082	FORWARD	58.09	CATTTACCCATCCAAAGCTG	258	(TA)9
	REVERSE	58.69	GATATTGGAGTCGGCTCTCC		
Brz0083	FORWARD	58.11	CATGATATTTGCCTGTCAAGG	251	(TC)5
	REVERSE	57.46	AGCACCGGTGATGTGAATA		
Brz0084	FORWARD	57.15	CTGCTTCAAATCTCGGATAAA	260	(TC)9
	REVERSE	58.03	ATCAAAACTGCTTTCGCAAC		
Brz0085	FORWARD	57.30	AATTCTGCCAATGATGCTTT	246	(AT)7
	REVERSE	58.53	AGCACATTTGTTCTCGCACT		
Brz0086	FORWARD	58.23	CGTGTGCAACAAAATTGAAA	240	(AG)9
	REVERSE	56.97	AAATCGCAAGGAAGTACTGG		
Brz0087	FORWARD	58.07	TTCCCCCACTACTCATCTCA	243	(GA)9
	REVERSE	58.45	AACAGCACACCGTAGCAAGT		
Brz0088	FORWARD	57.40	TTGTTCCAAAACTTGAATCTGA	255	(GA)5
	REVERSE	57.53	CCACTACAGCTCGACAATAGG		
Brz0089	FORWARD	58.47	CAAACCTATTCCACGGTCAA	249	(TC)13
	REVERSE	57.23	TGGACAATGCTATTCAAACG		
Brz0090	FORWARD	57.51	AAAGTCGCTGACACTATGATGA	260	(AC)7
	REVERSE	58.00	GCTTGATGACCTACCACCAC		
Brz0091	FORWARD	58.15	TCCGATCAGGGTCAAAGTTA	257	(AT)5
	REVERSE	56.94	CCAAGTACACATGCCCATTA		
Brz0092	FORWARD	58.07	TTGATCAGTGGGAGGTAGGA	251	(AT)6
	REVERSE	57.85	TGAAACTTGTCCCTTTTTCG		
Brz0093	FORWARD	56.54	CAACCAGCCTTAGTTAAATGG	259	(TA)5
	REVERSE	56.01	CCTGACTGGGCAGTAAGTTAT		
Brz0094	FORWARD	57.97	ATGATTTGATACGCCGTTGT	256	(AT)6
	REVERSE	57.78	CTTGGGACAAAGCCAAAGT		
Brz0095	FORWARD	57.60	TAACATGGCTGTTGTGGAAA	251	(CA)5
	REVERSE	58.00	ACTCTTCATCCGGTGGTGTA		
Brz0096	FORWARD	57.23	ACAAGTTAGCCTTGCGACTC	249	(TA)5
	REVERSE	57.94	CCAATTGTGGATGGCTTAAC		
Brz0097	FORWARD	58.32	TAATTTGTTCCACCCACAGG	244	(AT)8
	REVERSE	58.03	GTGACAGAGTTCGGGAGCTA		
Brz0098	FORWARD	58.28	AGCTTGCACATAGCAGAAGG	249	(AT)9
	REVERSE	57.67	TTTTTGTGGCACACAGGTAA		
Brz0099	FORWARD	58.75	TCGATCGGAGAACTGATGTC	245	(AT)9
	REVERSE	57.94	TGGATCGGACATACTCCTGT		
Brz0100	FORWARD	59.08	CCATCTGCAATTATTCAGGAAA	256	(AT)11

	REVERSE	57.64	GTTCTTGGTGCTTGACCATT		
Brz0101	FORWARD	58.32	TGCAGAAGCATCTTGCAGTA	254	(TA)9
	REVERSE	57.78	ATGCGCAGAAAATACAAACC		
Brz0102	FORWARD	58.79	AAAACTCGCCATGAGAAGGT	248	(TC)5
	REVERSE	58.39	TTTGTGATCGGCTTGCTTAT		
Brz0103	FORWARD	58.32	CGTGTATTCGTAAGGGCAAG	250	(TA)5
	REVERSE	57.91	AGGACCAATCATGTTGGAGA		
Brz0104	FORWARD	57.26	TAAGCCAATTAAGCCAAAGC	245	(AT)5
	REVERSE	58.83	GCGGTAACATTACCCGATTT		
Brz0105	FORWARD	59.09	CTGATCATTCCTGGTCAACG	250	(CT)7
	REVERSE	58.16	TGGCGGGATTTAAGTAACAA		
Brz0106	FORWARD	57.97	TGAACACACAGGTTCCATTTT	254	(TA)5
	REVERSE	57.64	GATGTCAACCAGCAAACCTT		
Brz0107	FORWARD	55.60	AGAGGAATTGACTTGGAAAAA	246	(GA)19
	REVERSE	57.85	GCATGCACGTAAATTTTCACT		
Brz0108	FORWARD	57.58	CCTGACTCTCAGGAAACTGC	256	(AT)7
	REVERSE	58.05	CGTCCAAAATCAGAAACCAC		
Brz0109	FORWARD	58.10	TTGAATTGTGGTCATTGCTG	254	(GA)11
	REVERSE	58.57	TGGCATGAAGGACCTATTTG		
Brz0110	FORWARD	57.97	CAAGCAGCAATTGGAAAGAT	254	(AT)11
	REVERSE	57.91	GGACAAGCTAGCCGAATGT		
Brz0111	FORWARD	58.00	GTGCTTCTGCATGGCTTAAT	251	(TC)9
	REVERSE	59.91	TATATGGAGGTGCCATGCAA		
Brz0112	FORWARD	58.16	CATGTTTGAACAACCTGCAA	243	(CT)7
	REVERSE	57.88	TCCATGTGTCTCTTCTGCAA		
Brz0113	FORWARD	58.00	AACAAGTAAGCTCTGCAGCAA	246	(CT)6
	REVERSE	57.24	TGAGTTGTACCAGTCGATGC		
Brz0114	FORWARD	57.92	GTGAGCGATGACTTGCCTAT	247	(GA)14
	REVERSE	58.89	AGCGACAGAAGGAAGGGATA		
Brz0115	FORWARD	57.96	AATTCATGATCGGAGCACAT	252	(AT)6
	REVERSE	57.67	TGAACAATGGCTTTGAATGA		
Brz0116	FORWARD	58.12	TCAAGAAATGGACTCCCAAA	250	(AG)16
	REVERSE	58.33	TCTAGGTCATGCAAGCCATT		
Brz0117	FORWARD	58.49	AGCTAAGGGGCTACTGTTGG	260	(TA)5
	REVERSE	57.72	CGCGATCTCCAAAATGTAAT		
Brz0118	FORWARD	58.30	AGGAGGTCCAAATCACCAAT	252	(CT)11
	REVERSE	57.69	CGTCAGCAATTCGTACCAC		
Brz0119	FORWARD	57.58	CAGATGACGTGAAGGGATTT	248	(TA)7
	REVERSE	57.89	ACCGACGAAATCATATTCCA		
Brz0120	FORWARD	57.45	CTGGTGATCTTACCCGTGAT	250	(CA)8
	REVERSE	58.57	GCACCCTCTTGTCACCATTA		

Brz0121	FORWARD	56.74	TGTCCTTCTCTCTCCTTGCT	247	(GA)7
	REVERSE	58.16	GCAATCATCCATTCATCCAT		
Brz0122	FORWARD	58.08	CATTGCTCCTCTCGCACTAT	253	(CA)6
	REVERSE	58.04	CTGCAGTTAGCAGGTTGGTT		
Brz0123	FORWARD	57.61	TCTCTAGGCCAACTCCTGAA	251	(TA)6
	REVERSE	57.97	GACAAGCCTAAAGCAATCCA		
Brz0124	FORWARD	57.60	AGGGACGCACACAATTTAAC	255	(CA)8
	REVERSE	58.01	GCTTTGCTTGACTTTGGTGT		
Brz0125	FORWARD	58.57	CAGCAAATGGGGTAAGATGA	254	(TA)5
	REVERSE	57.98	TCTCACACAAGCAGCAATGT		
Brz0126	FORWARD	55.70	AACAGAAGATTCCCTTCCAC	256	(TA)5
	REVERSE	58.20	GCCATCGTTGCCTTGTAA		
Brz0127	FORWARD	57.29	TCCGTTCTGATAATCCTTTGA	249	(TA)7
	REVERSE	58.28	GAGTGGGTTGTCTGCTTGAC		
Brz0128	FORWARD	58.05	AAGACGCATCAATTCTCAGC	256	(CT)9
	REVERSE	57.22	TTTCAAACCTGGTACCAAAAA		
Brz0129	FORWARD	57.93	TTGTACGTGCGTCTAGTGGA	254	(AT)6
	REVERSE	57.31	CGTTCTACCCGGTTTTGTAG		
Brz0130	FORWARD	57.99	TCCTTTCATGAACCCCTGTA	248	(CT)14
	REVERSE	58.33	CATCGCACGCTTATATGACA		
Brz0131	FORWARD	56.09	TGCAATGACATTAAATCAACC	260	(TC)7
	REVERSE	56.54	GCTGCAACACAAACAAAATAA		
Brz0132	FORWARD	58.04	CAGCGTTACAGAGGTTCGTT	251	(TC)9
	REVERSE	57.17	TTGCTAAACAAGCTGTTCCA		
Brz0133	FORWARD	57.44	TACGCGCCTACTGATGTATG	251	(GA)5
	REVERSE	58.59	CTTCTTCCTTTCCTCCGAGA		
Brz0134	FORWARD	58.00	AAGTGCAATATGAGCCGAAG	254	(CT)8
	REVERSE	57.47	GCAGTATTGCTGGTGTAAAGG		
Brz0135	FORWARD	58.36	ATGCCCAGAAGAGGAATAGC	253	(TA)7
	REVERSE	57.83	GAAGCATGTGTCAAGCAATG		
Brz0136	FORWARD	57.97	TTTTTACCCACCTTGGTTCA	251	(GA)9
	REVERSE	57.79	CTGCCAGAGGCAACTTTTAG		
Brz0137	FORWARD	58.17	GAAGCTTGAGCCAAATGAGA	242	(AT)9
	REVERSE	57.97	AAGTATGGCGTGGTGGATAA		
Brz0138	FORWARD	57.84	TACTCACATTGCCTTCGACA	250	(CT)7
	REVERSE	58.15	ACGTGCAGAAAAACCCATTA		
Brz0139	FORWARD	57.95	ACCATGCCGTGATAGTTTGT	243	(TC)9
	REVERSE	57.67	TCTGGAACCTAACCGAAATG		
Brz0140	FORWARD	58.29	GCAAGTGTGGTGGGATATGT	254	(AT)9
	REVERSE	57.47	CACATTTGACCATACGCAAC		
Brz0141	FORWARD	58.07	GCGCTTAGAATTCCTGATGA	253	(AG)8

	REVERSE	57.74	GAGGCCATCTAACCAAGTCA		
Brz0142	FORWARD	58.80	GCTGGGTTATGCTAATGCAA	253	(CT)15
	REVERSE	57.75	TCAAGCATGAACATTGAAACA		
Brz0143	FORWARD	57.86	GAGCTGTGGACAAATTTTGAA	243	(AT)7
	REVERSE	55.47	CTCCGAATATAATGCGGTTA		
Brz0144	FORWARD	58.33	CAGCGTGATGGAGGATTTAG	255	(CT)17
	REVERSE	57.97	CCAATGCTGTACTTCTCTGGA		
Brz0145	FORWARD	56.77	CAGGGGTGTACTTCTTCGTT	260	(GA)8
	REVERSE	58.20	ACCCATTTTCAGAGCACAAA		
Brz0146	FORWARD	57.82	TGTTGCTAGCTTTGCAGATG	260	(TA)6
	REVERSE	57.42	AACTTTTGCTGATGGAGCAT		
Brz0147	FORWARD	58.19	CTGAGGACGCTCCTACTGAA	244	(GA)14
	REVERSE	57.89	TTGATTTCAACACCCCAACT		
Brz0148	FORWARD	57.87	GCTCTTGACCTTGACGATGT	255	(TC)10
	REVERSE	57.24	TGCACTTGAGAGAGACGAAA		
Brz0149	FORWARD	58.26	GCAAGACCGCTGTTAGAGAA	245	(AT)11
	REVERSE	57.83	CTAACATGGACACCGCTCTT		
Brz0150	FORWARD	58.11	CAAGGAACAGAGTGGTGGTC	246	(CA)8
	REVERSE	58.02	ATACTGACCATGCCAAGGAA		
Brz0151	FORWARD	57.62	CGAGAGTACGAGGTTTGACTG	245	(TA)8
	REVERSE	57.71	CTGACCTAACCCCACTGAGA		
Brz0152	FORWARD	56.92	ATGCTGCACTTACTGGTTCA	246	(TC)11
	REVERSE	57.76	GGCTATCAATTCGAAGACCA		
Brz0153	FORWARD	58.02	AACACAAGGGAGAGGGAATC	251	(AC)16
	REVERSE	58.40	TGTTGGTCTTGCAGACAGTG		
Brz0154	FORWARD	57.37	GACAATAATGCATGTAGCTTGG	255	(CT)7
	REVERSE	58.48	TCCCCCTCTCTCTCTCTCAC		
Brz0155	FORWARD	57.84	CAGGTTCCAGGAGAGAAACA	252	(GA)10
	REVERSE	57.96	GCACCTCGTGTCTACGTCTT		
Brz0156	FORWARD	58.27	GCCATGATGTTTCATTGGTT	260	(AC)7
	REVERSE	58.41	TTTTGCACCTTTCATTGCTT		
Brz0157	FORWARD	58.08	AGTTGACCGCACATCAAAAT	246	(AT)10
	REVERSE	58.00	GTCGACTTGCAAAGGAAAAA		
Brz0158	FORWARD	58.13	GTCGCTGATCTGCAGAGATT	251	(AT)9
	REVERSE	58.29	CGCAGTCAATGTCGTCATAA		
Brz0159	FORWARD	59.26	GGATTATCCACGTGAAGATGG	258	(AT)8
	REVERSE	52.89	TCATAATCATAAAGCATGAAAA		
Brz0160	FORWARD	57.99	GAAATGTTGATGGGCTGAAC	253	(CT)6
	REVERSE	56.78	ACATCTAGCATCGTCGATCA		
Brz0161	FORWARD	58.17	GCTAGCGTGGAAACAAGAAA	293	(TC)9
	REVERSE	58.02	GAAATGCCCTGGTAGATGTG		

Brz0162	FORWARD	58.27	TGCCAAAAGAGGGTTGTTTA	292	(AT)6
	REVERSE	57.74	TTACCCATTTGACTCAGTTCG		
Brz0163	FORWARD	58.04	GGCATTTGCGACTTTTTAGA	305	(GA)9
	REVERSE	58.38	CACGGAGCTTCAAACTATGG		
Brz0164	FORWARD	58.42	TATTCCACTTCCCATGTTGC	298	(AT)6
	REVERSE	57.95	ACACCAACAAGTCCCAAAAA		
Brz0165	FORWARD	57.95	TTGACGGAGAACTCGTTAGG	301	(TA)7
	REVERSE	58.28	GCTAGCATCTTGCTTGTCGT		
Brz0166	FORWARD	57.35	TGCATCAAATTGTTCTTTCG	296	(AT)4
	REVERSE	58.75	TATGGCAATCTAGCCACGAC		
Brz0167	FORWARD	58.50	ATGAAACTCCAAAGGCCAAT	295	(AT)6
	REVERSE	58.65	ATGCAACATGCACAATCTGA		
Brz0168	FORWARD	58.34	TCCTTGTTTACACCCCAATTT	309	(TC)8
	REVERSE	57.90	GCTGGCATTTCCAACTTTTA		
Brz0169	FORWARD	57.68	TCTGGCAAAATTAAGTACAGCA	300	(TA)8
	REVERSE	57.83	AACCCAACACGACAAAATGT		
Brz0170	FORWARD	57.50	TTCAAATAGAGGCAGTTTCCA	299	(CT)6
	REVERSE	58.07	TGAATACGAACAAGCAAGCA		
Brz0171	FORWARD	57.33	TTGTCTCACTTGTGCACTCC	305	(AG)8
	REVERSE	57.75	GCTAGCAGGTAGCAAGATGG		
Brz0172	FORWARD	57.95	CGACATCACCTTTGCTTTCT	301	(CT)8
	REVERSE	58.13	CGCTGCTACTCAACAGGAAT		
Brz0173	FORWARD	56.73	ACCTGGACAGAGAGGTATGC	292	(AT)6
	REVERSE	58.54	TGCAATTGTCGATTGTGTGT		
Brz0174	FORWARD	57.79	AATTGTGTTTCATGGGCATT	316	(GA)5
	REVERSE	58.20	CACCTTCCGAATGAAAACAG		
Brz0175	FORWARD	58.13	TCACCAATCGTCTTGTTCCT	298	(AG)8
	REVERSE	58.55	ACCCAAACAAGCATTCCTTT		
Brz0176	FORWARD	58.32	CGAGAGATGATGAGGGAGTG	306	(AT)5
	REVERSE	57.47	TCCTGGATGAACATGTGAGA		
Brz0177	FORWARD	57.55	TGGAGTTGAGGCTTTAGGAA	304	(TC)7
	REVERSE	57.70	GTGTTTGGAAACCACTTGCT		
Brz0178	FORWARD	57.54	TGTCATGACCCTAATGCAAA	313	(TA)8
	REVERSE	58.49	AAACTCAGGGGGTTCTGTGT		
Brz0179	FORWARD	58.33	TTTCGGAGGAGGAGATTAGG	299	(TA)8
	REVERSE	58.27	TGGCCAACACCTTAGAAAAA		
Brz0180	FORWARD	58.41	CACACGGTCCATCTTGATTT	302	(CT)6
	REVERSE	58.16	TCCATAATGCATTGTCTTGAAA		
Brz0181	FORWARD	56.89	AGGGGGAGATGTTATTGTCA	296	(AT)5
	REVERSE	58.19	AAGCCGTTGCCAATTATGTA		
Brz0182	FORWARD	57.92	ACGTTATTGGACTTGGGTGA	317	(CT)9

	REVERSE	57.80	AGCCTGACCAAATTCTTGTG		
Brz0183	FORWARD	58.41	TTGAATCACTGGTGGGTAGG	295	(TA)5
	REVERSE	58.01	GAATTTTGAGTCCAACCCTGT		
Brz0184	FORWARD	58.16	CAGGATACATAGCCTGGGACT	302	(AT)5
	REVERSE	58.05	TGAATCGGAAAACCTTGTGT		
Brz0185	FORWARD	57.93	AGGTATTGCAAATACGCCTCT	290	(CA)5
	REVERSE	57.94	CTGGAGGGTATCCATGTGAG		
Brz0186	FORWARD	58.30	CAAGCCCAACTACCCTGAG	298	(AC)10
	REVERSE	57.97	GTGGATACAAAGCCATACCG		
Brz0187	FORWARD	57.99	GTTTGTGGCCTCATTCATTC	297	(AT)7
	REVERSE	58.11	TTTTTCAAAATCCGGTGAAA		
Brz0188	FORWARD	58.19	AGTGTAGCTTGTGCCGAAAG	298	(AT)5
	REVERSE	57.88	TAGCGAAAGCGAGTCTTCAT		
Brz0189	FORWARD	58.24	AGCCAAGCAGGCTTCTTTAT	298	(AT)13
	REVERSE	57.61	GTGGATCCAATCCAATTGTC		
Brz0190	FORWARD	58.15	CGCGGGTAGCTAGTGTCTT	290	(AT)5
	REVERSE	58.20	GCCCTAAGAATTTGTCGTCAT		
Brz0191	FORWARD	57.19	CAGTTTTCTAACTGCTCCATGA	305	(AT)5
	REVERSE	58.11	CAGCCCAATAATGCAAAATC		
Brz0192	FORWARD	57.96	GGTCGCTGTCATTATCCATC	297	(AT)6
	REVERSE	57.94	AAATGTGTGCACATGAGTGG		
Brz0193	FORWARD	58.29	ATATACTCCCAGCTTTCACACG	290	(AC)6
	REVERSE	57.98	GAGGGCAATAGCCTCTCAA		
Brz0194	FORWARD	58.17	TCCCAGGACAAGCTATGGTA	305	(CT)9
	REVERSE	57.48	TCCACACCTCTATCCCAGTC		
Brz0195	FORWARD	57.45	AACGTTGTGGAAGAAGTGCT	291	(AT)5
	REVERSE	57.92	AATTGTTCCAACGACGACAT		
Brz0196	FORWARD	58.50	CTGCAAAGGCAGATTGACA	301	(AT)4
	REVERSE	54.79	CGCCATGTCATATATCATAAAC		
Brz0197	FORWARD	57.29	TCTTTTGCCAGAAAAGTTCAG	294	(AT)6
	REVERSE	58.33	TGTCCATGCAAATACATCCA		
Brz0198	FORWARD	58.00	GTCCATCCTCGTTAGTGGTG	298	(AT)7
	REVERSE	58.38	GCCTCCAAAAGAGGGTTAGA		
Brz0199	FORWARD	58.73	GAGCAACATGGTGCATTCTT	300	(TC)7
	REVERSE	58.05	GCATGTCAGTCACGGTATGA		
Brz0200	FORWARD	57.90	TTTCACGAGCACAGTTACCA	297	(AT)6
	REVERSE	58.04	TTCAGATGCCACATTTGATG		
Brz0201	FORWARD	58.17	CTCTATGTTGTCCGGATTGC	290	(GA)8
	REVERSE	57.49	GGTGGCCCACAAATAAAGTA		
Brz0202	FORWARD	58.22	GGAGAGACAACAGCATGGAC	317	(AT)11
	REVERSE	57.91	TGAGTGACAAACAATGGGATT		

Brz0203	FORWARD	57.69	CGCTTGAGAAGCTAGCAAGT	301	(GA)8
	REVERSE	57.93	TAGCCTTTTGCATGGGTTAG		
Brz0204	FORWARD	58.16	TCGTCCTCGACAACCTGTAT	291	(TA)4
	REVERSE	56.63	AAATCGTGTATGCATGTTCAA		
Brz0205	FORWARD	57.63	AAATGGGAACGTTAATGCAG	294	(TA)5
	REVERSE	58.20	CATCCTCGTTTCCACTTTTG		
Brz0206	FORWARD	58.23	GAAGTGGCAAGACACACACA	297	(TC)15
	REVERSE	58.31	TGAGCTTTTCGTCTCTCCTG		
Brz0207	FORWARD	56.93	CCTCGAGGAGATACAAGGAA	299	(TA)6
	REVERSE	57.82	TAGGACCGTTCTTCATTTCG		
Brz0208	FORWARD	57.42	GGTGACCTGGTCTATGGAAA	303	(TC)5
	REVERSE	57.92	TTTTCTGGGTGAATTGGGTA		
Brz0209	FORWARD	58.35	CGCAAGAAAACAGAATGACC	303	(AT)4
	REVERSE	57.48	TCATGATCCAGGCATTACAA		
Brz0210	FORWARD	58.60	CCGTGGCATAGAATATCGAA	314	(TA)4
	REVERSE	57.89	AGAGCCATACCCTAGACTCCA		
Brz0211	FORWARD	57.87	TTTATATCTTGGCGGACAGC	295	(TA)7
	REVERSE	58.62	AAGGGTTCCTTTCTGAACCA		
Brz0212	FORWARD	57.73	ACTCATTTTCACACGCACAA	301	(CA)5
	REVERSE	58.26	CGAAGAATTGCAGCAGAAGT		
Brz0213	FORWARD	57.01	TGAAGCCCTTTCTAAATGATG	296	(CA)7
	REVERSE	57.74	GAACTAGGAAGCCATGGACA		
Brz0214	FORWARD	57.58	TCTGGTGTCTCTTTGCTCCT	309	(AT)8
	REVERSE	57.27	TCCATGGTACCTGAATGACA		
Brz0215	FORWARD	58.20	TTAACCTGCAGCAAGTAGCC	300	(AG)8
	REVERSE	57.66	TGCAACAATACCCAGCATAA		
Brz0216	FORWARD	58.19	ACGAGAAGCTCGACAATCTG	300	(AG)9
	REVERSE	58.01	AGGTCAGCGGTTCTCCTAGT		
Brz0217	FORWARD	57.80	GGTCCCTGTGCTCAGTTTTA	302	(TA)5
	REVERSE	58.07	ATAGGTAGCCCGTCAAAACC		
Brz0218	FORWARD	57.85	CTTTGCATATGGTTGCTCCT	300	(AT)6
	REVERSE	57.65	CATTGGGAGAGAAGATCCAA		
Brz0219	FORWARD	57.76	GCAGTTCTTGCTTTTTCAGG	302	(AT)4
	REVERSE	55.61	TCTCCTTATGCAAGGCTTC		
Brz0220	FORWARD	57.14	TGCGAATGATATAACAGAAAGC	301	(AC)6
	REVERSE	58.03	TGCCATAAAATTTTGCCATT		
Brz0221	FORWARD	58.44	TGCTAAAAACGCCATAAGGA	295	(TA)4
	REVERSE	57.90	TGTAGATCGATGTGAACTTTGC		
Brz0222	FORWARD	58.48	AAATGATGCCAAAAATGACG	348	(CT)6
	REVERSE	57.90	CGTTGTTTGCATCTGTCAAG		
Brz0223	FORWARD	58.67	GGTCACAATTTGGGTCACAC	294	(AG)8

	REVERSE	58.09	TCGCAAAAATTCTTCTGGAG		
Brz0224	FORWARD	58.05	TGTGAGCAAAATTGAAAGCA	304	(CA)7
	REVERSE	57.45	TTTGGTGTTTTGCCTCTTCT		
Brz0225	FORWARD	58.44	TTTTGCTCGCACAATAGGTT	321	(AT)8
	REVERSE	58.05	TCGACTTGCGAGCATATACA		
Brz0226	FORWARD	57.92	GCATGGTGCGATAGTCTTCT	306	(AG)5
	REVERSE	57.57	TCTGGCTGACTGTGACTCTG		
Brz0227	FORWARD	57.91	AGACCTACCGCTCCGACTAT	291	(TA)6
	REVERSE	57.75	GTTCACCACATCAACACCAA		
Brz0228	FORWARD	57.85	CCACCTCATCATTTGTGTCA	307	(AT)7
	REVERSE	57.92	CTATGGGGTCCCATCTCTTT		
Brz0229	FORWARD	57.15	GCTGTGTTTTGGATTGTTCA	303	(TA)5
	REVERSE	58.87	CCATATAACGACGCACCATC		
Brz0230	FORWARD	58.07	CCGTTGGAATTCTAATTAACCA	292	(AT)6
	REVERSE	58.06	CCGAGGGTCTTGATTTGTC		
Brz0231	FORWARD	57.13	TGCAAAAGAACGAACAAAGA	296	(AT)5
	REVERSE	57.84	ACCACTTTGGTCACGATGAT		
Brz0232	FORWARD	57.76	CGGCAGTTTTAACATGACCT	299	(AT)6
	REVERSE	57.64	GCACCATGATTCCTCTGACT		
Brz0233	FORWARD	55.41	TTTTTATTTGCATTAGCCTGA	300	(AT)8
	REVERSE	57.73	CAAAATCCGAAACCTTTTCA		
Brz0234	FORWARD	57.64	CACGGGATCAATTGCTCTAT	304	(TC)9
	REVERSE	57.94	CAACAACGGTCATCATTGAA		
Brz0235	FORWARD	58.33	CACACTCACACACGGAGAGA	298	(TC)9
	REVERSE	57.81	CATCCAGAGCCTGATGAAGT		
Brz0236	FORWARD	58.38	CCTTGGTGCTTCATTCTACG	303	(CA)9
	REVERSE	58.41	GCTCTTGAGCTCACTCCTGA		
Brz0237	FORWARD	58.08	TTGTCCATGCCTAGGTTGTT	310	(AC)7
	REVERSE	58.06	AATGGATGGGTGAAATGATG		
Brz0238	FORWARD	58.03	TACACACCATCACGCAAAAC	303	(CA)5
	REVERSE	58.35	AATCCACGACCACTGATGAC		
Brz0239	FORWARD	58.23	AGGCGTAAGACAATTGGTGA	308	(AG)8
	REVERSE	58.95	CGCAGGTTGCTACAACATCT		
Brz0240	FORWARD	58.34	AGTGCTCGAGGCATACACAT	310	(TA)5
	REVERSE	55.07	AAATCCCAGAGTACATAAATCG		
Brz3001	FORWARD	57.67	AAAGATGACATTGCCGTTTC	149	(TGA)6
	REVERSE	58.02	TTCAACTCATCGTCATGTGG		
Brz3002	FORWARD	58.75	GCTGGAATCAGAATCGATGA	155	(TTG)9
	REVERSE	58.03	GAACTGCAGTGGCTGATCTT		
Brz3003	FORWARD	57.27	GTTTCAGGAGGCATACAAGG	160	(AAT)7
	REVERSE	57.67	CAAGGCAGGAAGGTACACAT		

Brz3004	FORWARD	58.28	TCATTCTGTGTGCGTGGTAG	154	(TTG)6
	REVERSE	57.26	AGAAACTTGCATCACCGATT		
Brz3005	FORWARD	57.55	CAGAGGGTTAATGCACCAAT	143	(TGA)6
	REVERSE	57.73	GGAGAAGCATCCAAAAATGA		
Brz3006	FORWARD	54.78	TTGATGCTTTATCACATTGC	148	(TTC)3
	REVERSE	58.19	CAGATTTTAGGCTGTGAAGCA		
Brz3007	FORWARD	57.64	GGGGTAATGTACCCAGGTTT	157	(GAA)3
	REVERSE	57.82	ATTGCGAGAAATTGACAAGG		
Brz3008	FORWARD	59.12	TAGTTTCAGAGGGGGAATCG	154	(TTC)3
	REVERSE	57.22	TGTGCCAAAATAACAGATGC		
Brz3009	FORWARD	57.42	AGACTCTGTGCGGGAAATTA	151	(AAT)10
	REVERSE	57.08	ACTTCGCTTGTCCTACTTGG		
Brz3010	FORWARD	57.85	AGGACAGTGTGTCGGAGAAG	148	(TCA)4
	REVERSE	57.20	GCAAGTTCCTTTCAAGCAGT		
Brz3011	FORWARD	58.31	GACTGGGGATTTTCCTATGG	248	(AAG)3
	REVERSE	58.11	AAAAAGAATGGATCCGAAGG		
Brz3012	FORWARD	57.59	GGCTTGCTGGAGAATCTTAAT	252	(GAA)4
	REVERSE	58.02	AATCCGCTTTTCTCGATCTT		
Brz3013	FORWARD	58.37	GCAGCAGTACCTTGACCAAC	252	(TAA)4
	REVERSE	57.82	TCGAAGTAATTCGGAGGATTT		
Brz3014	FORWARD	58.23	CAGGAACGATGGAGAAGATG	248	(GAA)10
	REVERSE	57.71	GGAAGAGATTCAAACCGTGA		
Brz3015	FORWARD	57.36	GGTAAGTGGATGATGGAGGA	239	(TCA)5
	REVERSE	57.95	GAGTGCCAACAAAGAGCAAT		
Brz3016	FORWARD	58.33	TAACCGCCCTGACAGAGATA	246	(AAG)6
	REVERSE	57.53	TTGAAATCTGCTATGCAAGGT		
Brz3017	FORWARD	57.99	TGGTGAAGGTTGGGATTCTA	264	(GAT)4
	REVERSE	57.86	CCTTTCTTGCCAAACACACT		
Brz3018	FORWARD	58.27	AGGTTCAGTGCGAGCTGTAG	250	(AAG)5
	REVERSE	58.09	GTGTGGCGTAGGTAGTGGTC		
Brz3019	FORWARD	57.02	TATGGTGAAAGTGTGCAACC	242	(AAG)4
	REVERSE	58.13	CGCCTAGAACTCAGCAACAT		
Brz3020	FORWARD	58.62	GCTCCTGGCCCTACACAT	232	(GAA)3
	REVERSE	51.05	CAACTTATTACAAGATGGAAAC		
Brz4001	FORWARD	58.21	CATTTTCGAGACGGATTTTG	144	(ATTT)6
	REVERSE	58.09	GGCTGGTATTTTCAATGCAC		
Brz4002	FORWARD	57.56	AACATGTCAAAGGAACAGTGG	154	(ATTT)6
	REVERSE	58.34	CCACACAGCAAACAATAGCA		
Brz4003	FORWARD	57.74	TGAAGTCCTCAATGATGCAA	147	(CATG)7
	REVERSE	57.58	GCTGGTTTAAGCATCCAGAG		
Brz4004	FORWARD	57.85	AGCAGGGAAGGTCAATCTTT	150	(AAAG)7

	REVERSE	57.60	CGAGCTAATTTCCTTGCATC		
Brz4005	FORWARD	58.23	TCGTGGTCTGTGTTTGAGTG	151	(TAAA)3
	REVERSE	58.75	ACATGCCTGGAGCTATTGTG		
Brz4006	FORWARD	57.69	TCCACACACACTTGTCTTTCA	159	(TTTC)3
	REVERSE	57.75	TAAATCCGCTTATGGCATTC		
Brz4007	FORWARD	57.98	TTTGCAAAAGAAAAAGATGGTT	158	(TAAA)3
	REVERSE	56.89	TCCTTGGTCTGAGCAAATATC		
Brz4008	FORWARD	57.97	TGTCGCATTCTCAAAAGACA	152	(AAAG)4
	REVERSE	56.81	ATATGCAAGTGTGTGGATGG		
Brz4009	FORWARD	58.44	AACGAAGCTAATTTGCCACA	140	(ATTT)4
	REVERSE	57.22	TTTTTCTTGCCCAGGTTG		
Brz4010	FORWARD	57.93	TCAGGGTGAAGGGAATATGA	159	(TAAA)4
	REVERSE	58.08	TACCTGCTGTTGGACCAAAT		
Brz3021	FORWARD	57.01	AGAGAATAACTCCCCGAAAAA	74	(AAG)8
	REVERSE	58.57	CAATGGCTTCAGGAATGGTA		
Brz3022	FORWARD	57.51	GTGGGAAATTTTGTGCTGAT	79	(ATT)9
	REVERSE	58.53	AATGATGCATTAGGGCCTTT		
Brz3023	FORWARD	58.75	TAGCAGGTAGTCGGTGGATG	73	(AAT)3
	REVERSE	58.28	AAAAACCTATTCCCCACCAA		
Brz3024	FORWARD	55.41	AGGTATACCAAGCAAGCTCTC	71	(TTC)3
	REVERSE	57.97	AGATCAGGAGCATGAACAGC		
Brz3025	FORWARD	58.19	ACCGGACTCTACTCCCACTC	80	(TCA)3
	REVERSE	58.19	CGATAGGGGCGGTAGTATCT		
Brz3026	FORWARD	50.66	TTAGGTTGTTGCTACTCTACTT	98	(ATT)4
	REVERSE	53.50	TTTGTTTTGGCTATATTCCTT		
Brz3027	FORWARD	58.16	GCGTAGTCAACACCATCTCC	78	(TCA)4
	REVERSE	58.08	TGGAAAAGAGAAGCAACCTG		
Brz3028	FORWARD	57.45	GAGAACGAATCCTCTGTTGC	82	(AAG)3
	REVERSE	56.66	GTCACTGACTGGTTTTGACG		
Brz3029	FORWARD	57.99	GCGCAAGAAAACTAAAGAGTACA	100	(TAT)3
	REVERSE	55.46	CTCCCTTCGCAAATAATAATAA		
Brz3030	FORWARD	58.23	AATTGTGGCATGCTGTTTCT	97	(TTC)5
	REVERSE	59.05	GCACGGGAAAAATAGGAAAA		
Brz3031	FORWARD	58.26	AGGAGGAGAGGTGGAATCTG	74	(AAG)7
	REVERSE	58.99	CCGCCTTTTTCTCCTTCTC		
Brz3032	FORWARD	58.01	ACAGAGGAGGCTGACTGTTG	70	(TGA)5
	REVERSE	57.14	CCCAGTTGCATTGCATATAA		
Brz3033	FORWARD	58.04	TAGTCTCTGCAGCGCTTTG	70	(CAA)7
	REVERSE	57.16	GTGTCTGCAGCTGTACCTTTT		
Brz3034	FORWARD	58.37	CGTGCTTACACGGAGATGA	79	(TGA)3
	REVERSE	57.92	TGAAGGTCGACTTTTTGAGG		

Brz3035	FORWARD	57.73	TCAGCAAGAAGCCTAATCGT	93	(CAA)5
	REVERSE	57.55	TTCATGTAGCCTACCCCAAC		
Brz3036	FORWARD	57.28	TGAGAAATCTCAAAATGATCCA	91	(TTC)3
	REVERSE	58.65	AACAAGAACGAATCGGACCT		
Brz3037	FORWARD	55.17	AAAAAGGACCACCCACTAAT	83	(GAA)7
	REVERSE	57.03	AGATCGATTCTTTGCCTTTG		
Brz3038	FORWARD	57.58	TCAGGAAGGGGTGTGAATAG	62	(GCC)3
	REVERSE	58.86	TTGTTGTTGTTGTGGTGGTG		
Brz3039	FORWARD	58.07	CAGTCCTCATGGATTTCTGG	98	(TGC)3
	REVERSE	58.00	TCTGGCAACTTTGAGAGTCC		
Brz3040	FORWARD	54.88	AATCTATTGCATGGTGTAGTCA	75	(CAA)6
	REVERSE	57.45	AAACTTGCTTGGGAGTGAAA		
Brz3041	FORWARD	57.45	CTGCAACATAAAACCTGCAA	75	(TCA)6
	REVERSE	56.68	CCTCAACAAGTCTCTTTGTTTG		
Brz3042	FORWARD	56.75	ATCCTAGCCTTAACGCAAAA	85	(AAT)7
	REVERSE	58.62	TACAAGCTTGGGATTGGTTG		
Brz3043	FORWARD	57.81	CGGCCAACAACTTAACAACT	89	(CTT)4
	REVERSE	57.73	TTTGGGAAACAGTTCTACCG		
Brz3044	FORWARD	57.81	TTGTCCAACAACTTACATGAGC	74	(AAT)5
	REVERSE	57.81	CCCATTTAGCATACCTTAGGC		
Brz3045	FORWARD	57.57	GGTTGTTGTTGCACAGATTG	86	(AGT)6
	REVERSE	57.23	CTTGAAGGAAATCAAGCTGAA		
Brz3046	FORWARD	50.28	GGAATATATCTCCTACCATTATAC	94	(AAT)6
	REVERSE	57.89	GAAATCGCACCCATTATGAC		
Brz3047	FORWARD	57.36	ACGCACAATCTCGTAGGTTT	94	(CAA)6
	REVERSE	57.44	TCTCAACATATCACGGACCA		
Brz3048	FORWARD	57.52	GGCAACCGGACAAATTATTA	79	(AAT)6
	REVERSE	57.99	TCTTTGGGGAGTGGAACATA		
Brz3049	FORWARD	57.65	GCTCTTCCTCACCTCCTTCT	78	(CTT)7
	REVERSE	58.68	GGTACAAGAAGCTGCAGTCG		
Brz3050	FORWARD	54.89	TTGATCTGGAAAACATGACA	95	(ATT)9
	REVERSE	56.67	TCACAGTGATCTGAACAGTAAAAA		
Brz3051	FORWARD	58.04	CCATTGCCAATATCCTCTTG	79	(TCA)6
	REVERSE	57.87	CCCATAGTTGGTCAGATTGG		
Brz3052	FORWARD	57.49	CGATAACTCGAAGCATAGGG	74	(AAT)5
	REVERSE	58.14	CAATATAGGGCCAGCAAATG		
Brz3053	FORWARD	54.92	TGCTTAATATCCCTTTTGATTT	99	(ATT)6
	REVERSE	58.22	CACACATGATCTCCGCTGTA		
Brz3054	FORWARD	57.78	GTCAGTTGACACTGGTGCTG	72	(CTT)8
	REVERSE	54.14	ATGAATTCTCATTCCAGAGG		
Brz3055	FORWARD	57.43	GTCGGCCTTGTTCTTCTTC	69	(AAC)6

	REVERSE	58.25	GGGTGTGTCTCAAATGTGGT		
Brz3056	FORWARD	58.37	CTTGTAGCCACACGAACTCC	80	(GTT)6
	REVERSE	58.12	CGCAGATCAGGGTAATCATC		
Brz3057	FORWARD	57.43	AAATGTATAGCCCGCTTTGA	98	(AAT)6
	REVERSE	58.16	TGGACCACAGTTAGCAGGAT		
Brz3058	FORWARD	58.68	GTGGGATCAGACGAGGAGA	80	(GAA)6
	REVERSE	58.74	CGCTTCCGCTTCTTCTTATT		
Brz3059	FORWARD	58.00	CCCTTCGTCACTACCATCAC	93	(GAA)9
	REVERSE	59.69	CAATGCAAATGCAGGTGTG		
Brz3060	FORWARD	58.75	CCTGCTTGCATGTAATTATTTTG	99	(CAA)7
	REVERSE	57.67	CAATTGCAGAGGGAAACATT		
Brz3061	FORWARD	55.89	GACAGTCTCCCTAGGTACAACTAA	100	(CTT)6
	REVERSE	57.93	GATGGGCAGGTGACAAAA		
Brz3062	FORWARD	58.39	TAAGGGGGTGCTTAATCCAT	68	(AAT)6
	REVERSE	58.29	TGTCCAAACATCCAATGTGA		
Brz3063	FORWARD	58.38	TGAGGTAAGCTTTGCCACAT	99	(AAG)6
	REVERSE	58.20	TCGATCGTTGGTGTTCTTTT		
Brz3064	FORWARD	56.12	TTCTCTTTTTAGTTAAACGTGGTC	66	(TTA)6
	REVERSE	56.59	TTATGCTCATGAAAGATGCAG		
Brz3065	FORWARD	57.38	ATGTTGCTCACTCTCGGTTT	82	(CTT)3
	REVERSE	58.23	ATGCATTTGGCACTGACTTT		
Brz3066	FORWARD	57.68	GGTTGATTCCATTGTTGACC	94	(GTT)8
	REVERSE	58.16	GCCTCGAGACTTGTGAAGAA		
Brz3067	FORWARD	58.35	CAGAGAAGCTGCGATTCCTA	89	(AAG)6
	REVERSE	57.92	CGGCGTCTCTCACTATGATT		
Brz3068	FORWARD	57.16	CCATAACAGATGAAACAACAGG	97	(TCA)4
	REVERSE	59.39	GCTGCCGCTGGTAAAGTAAT		
Brz3069	FORWARD	58.05	TGCAGACGTGAAGAGAATCA	66	(ATT)4
	REVERSE	53.49	TCGAATAATTGGAAACAAAA		
Brz3070	FORWARD	58.19	TAATTCCTTCCCTCCCTCTG	96	(TGA)6
	REVERSE	57.45	CTTGTCCACTTTTATCATGCAG		
Brz3071	FORWARD	57.98	TGAAATTGAAAAGTGAATTCTTGA	87	(AAT)5
	REVERSE	57.35	CATGCATTCAAGACTTTTTCC		
Brz3072	FORWARD	59.02	CCAAATTGTGCGAAAACATC	86	(CAT)9
	REVERSE	55.92	CAACAATATTTTGCCCCTAGT		
Brz3073	FORWARD	51.13	GATGCATGTTTATCATCTATCT	97	(AAT)3
	REVERSE	53.59	CGACTATTAATCAATAAATTAGGG		
Brz3074	FORWARD	57.78	TGACCCGTCTTCTTCTACTAGC	98	(ATT)7
	REVERSE	50.40	TCTTATATTATCTCAACGTAAAAA		
Brz3075	FORWARD	58.37	ACACAGCAGCAGAAATGGTT	70	(AAG)6
	REVERSE	56.83	GCTCAACGTGCTAATTGCTA		

Brz3076	FORWARD	57.32	TCTCCTTCTGCTTCTTCGTC	80	(TTC)5
	REVERSE	57.09	GATCCAATCGGAACGAATAG		
Brz3077	FORWARD	57.76	TGATTTAGTTATTGTTCCCTTTCC	67	(CTT)7
	REVERSE	57.49	TTGATCTAATTCATTCGCAAAA		
Brz3078	FORWARD	56.96	TTGTTCTTGTTATTGTTGTTGTTG	98	(TTA)4
	REVERSE	57.26	CTTGGCACAACATTTTTCAA		
Brz3079	FORWARD	59.00	TTGGTTTTGGACAACCTGAA	99	(AAG)7
	REVERSE	55.48	TTCTATTCACTCGTCCTGTTG		
Brz3080	FORWARD	57.18	ACGGACATTATGCCTCCTTA	79	(CAA)4
	REVERSE	55.38	TTACGTCCCAACTACTACCG		
Brz3081	FORWARD	58.09	CCAATTTGCCTATCACAAGG	90	(ACT)5
	REVERSE	58.19	GCAGACGGAGACGAAGAGTA		
Brz3082	FORWARD	56.60	TTTTTGAACGTACATCATCCA	88	(ATT)3
	REVERSE	58.21	CCCATATCCTTCCTCTTGCT		
Brz3083	FORWARD	58.21	GAAGCATCCAAAAATGCAAC	99	(TCA)5
	REVERSE	58.72	CCACCTCAAGCACTTGGATA		
Brz3084	FORWARD	57.88	CAAAAGGTGAAGCCAAGGTA	72	(TGA)3
	REVERSE	58.02	CATCAAGGTCATCATCATCAAG		
Brz3085	FORWARD	57.89	GAGGACCTGTGGATGACAAC	93	(GAA)5
	REVERSE	58.32	AAGTTTCAGAGGGGAGATGC		
Brz3086	FORWARD	57.41	TGGGTTTCAATTCTAGCCTATC	98	(GTT)8
	REVERSE	57.68	AACGTCTTTCACTCAAAGCAA		
Brz3087	FORWARD	56.48	AACAACTCGATGATTGGTCA	86	(CTT)3
	REVERSE	57.86	ATAAGGGTATGCCTCGCTTT		
Brz3088	FORWARD	57.47	CTGACCCCAAGGATTGACT	66	(TCA)4
	REVERSE	57.64	GACTTGAACAAGCCACCAAT		
Brz3089	FORWARD	57.61	TAACAGGAGAAGCCAGAGGA	100	(AAC)6
	REVERSE	58.64	GCCCCAATTATTGATATTCCA		
Brz3090	FORWARD	58.93	CGGTTTAATAACGAACCGTGT	70	(AAC)3
	REVERSE	58.37	CGATAGGAAAAGAATTCGATAGG		
Brz3091	FORWARD	57.62	GAAGTGACTCCATTCCCAGA	73	(TGA)8
	REVERSE	58.86	CCTGAGCATTTTCCTCTTCC		
Brz3092	FORWARD	57.45	GGCTGACAGGGAATGTGTAT	89	(GAA)3
	REVERSE	58.08	AGTTCCGCGAGAAGGTAGTT		
Brz3093	FORWARD	58.09	CGGTGCTATGTTTGATTTCC	67	(TGA)4
	REVERSE	58.03	AACCTTGGTTGGGTCCTTAG		
Brz3094	FORWARD	55.46	ACTCCCCTCACTTTTCCTAA	99	(AAT)5
	REVERSE	55.34	AGGGTGAGAATCTATTTGTTTTT		
Brz3095	FORWARD	58.94	CACACCAGGAAGTGACCATC	99	(TTC)3
	REVERSE	58.13	AGAGCATGGCGAAGTAGTTG		
Brz3096	FORWARD	57.09	AGATGAGGACATGCCAGAAT	77	(TGA)5

	REVERSE	57.32	CATTCATCAAAGCCATAGCA		
Brz3097	FORWARD	57.59	AGCAATCAGTCTGTGAGGAGA	90	(CAA)4
	REVERSE	57.00	CCGGAAAGTAGCTATGTCGT		
Brz3098	FORWARD	58.37	AGCAAGTACGAGGTTGACCA	85	(AAT)5
	REVERSE	55.56	AGGAAATCCTAACGAGCAAT		
Brz3099	FORWARD	57.55	AAGGGGGAGTGTTGAGAAAT	100	(CAA)4
	REVERSE	58.59	CTTGACACGGTTGAGAGAGC		
Brz3100	FORWARD	57.87	TCCCATAGGTCTTGGTGGTA	69	(TTG)3
	REVERSE	58.20	TTTTTGGCTGTGGTCATTCT		
Brz3101	FORWARD	50.72	TTAATAGGTGGTGACTTCAA	97	(AAT)4
	REVERSE	53.87	TGATAGCATTAAATAGGAATGG		
Brz3102	FORWARD	57.73	GTGTTTGTTCGTGCATCTTG	100	(AAT)5
	REVERSE	58.83	ATGGATGCATGCGACATTAT		
Brz3103	FORWARD	57.75	GACCGATACGCTAGAAGCAG	68	(AAG)4
	REVERSE	58.45	GTGCTTCGATGATTGCTTTC		
Brz3104	FORWARD	57.86	GAACCTGAAAATGAGCCAGA	88	(GAA)3
	REVERSE	58.03	CATCATCTGCAGCTAACGTG		
Brz3105	FORWARD	57.70	TGGCAAGCTCCTACAGTTCT	83	(TAA)3
	REVERSE	57.52	GATCGTCTTTCATTTTGTTAGTGA		
Brz3106	FORWARD	58.22	CAAGGTTACCAGGTCACAGC	63	(TTC)3
	REVERSE	58.24	CTGATTTCTTTCGGGAAGGT		
Brz3107	FORWARD	57.76	GCACCTTCTTTACTGGCTTCT	95	(GAA)3
	REVERSE	57.87	TTAGTGAACTCCCTTGTGTCG		
Brz3108	FORWARD	58.72	CCCCCAGGTCATCAATTAGT	86	(AGT)3
	REVERSE	57.82	TCTTTGGTCTGCCCATTAAG		
Brz3109	FORWARD	58.30	ACAAGGTCAGGATCCCATTT	93	(TAA)3
	REVERSE	57.58	CCTGGGAACAACATAGAGGA		
Brz3110	FORWARD	55.57	AGTGAGACTTCTTATTACCCTTCA	88	(TGA)5
	REVERSE	56.93	TTTCCTTAACATCTTCATCACG		
Brz3111	FORWARD	58.63	TGCATTTACCGAATTTCCTG	99	(ACT)3
	REVERSE	57.09	GCCGTGTATTAGTAAGGATAGGTT		
Brz3112	FORWARD	57.81	GCCAGCAGCTTAGAACCTATT	93	(AAG)3
	REVERSE	57.42	CAGCCGCCTTACACCTATAC		
Brz3113	FORWARD	51.85	TTTCAGGGCTACACTAAAAT	90	(ATT)5
	REVERSE	58.33	CTTGGCCGAAGGATTAAGTT		
Brz3114	FORWARD	57.74	TGAAAGGCAGATACACAAACC	98	(ATT)5
	REVERSE	57.23	AAGGGTCAGCATAATTTGTCA		
Brz3115	FORWARD	55.98	CAGCCTCCCTAGATTCATTT	100	(TTA)5
	REVERSE	55.09	CATAGCCTCCAACATAAAGC		
Brz3116	FORWARD	55.08	TGAAGTCAAAGTAAGAAGTGTTGT	77	(ATT)5
	REVERSE	57.61	GGTGCAAACCCCTACTGTTA		

Brz3117	FORWARD	58.20	CTACCCTACCCCAACTTGCT	75	(TTG)6
	REVERSE	57.66	CGTTAAAATGCATGACGATG		
Brz3118	FORWARD	58.99	GCAACAGAAGCCCTCAAAA	74	(TAA)5
	REVERSE	57.86	TTCATTCGGATTTGTTGGAT		
Brz3119	FORWARD	54.75	CTTTAATAAGACAGGACAGACAAA	72	(AAC)5
	REVERSE	57.11	TGTTCTGTTTTTCCCAAACAT		
Brz3120	FORWARD	57.96	CTATCAACATTCAACCCGATG	84	(TCA)5
	REVERSE	58.81	GGGTTCGTTGTAGGGTTTGT		
Brz3121	FORWARD	59.71	TGTACGACGTCGCTGCTC	82	(TTG)6
	REVERSE	57.39	TGAAGACGAGGAGGAGGATA		
Brz3122	FORWARD	57.97	CGCCTAAGACCATTCTGAAA	76	(TCA)4
	REVERSE	56.19	TGTCATGTTCACATATTCTTGC		
Brz3123	FORWARD	58.20	CAGCAGATAATCCCCAAACTT	76	(AAT)7
	REVERSE	57.70	CATAGCGTGTTGGTTGTTTG		
Brz3124	FORWARD	57.62	AGGAGCTTGAGCAGATAGCTT	100	(TCA)5
	REVERSE	53.99	TTGTCTATCATAAAGTTGTTGTTG		
Brz3125	FORWARD	56.82	ATGTTCAGGCTTCCTTTCAC	89	(ATT)6
	REVERSE	61.19	TGCTACGGGTGGCAAAAA		
Brz3126	FORWARD	56.89	TGAATTTCGAGTTGTGTGCT	66	(TTC)3
	REVERSE	57.39	GGAACCACAGCAGTCTAAGG		
Brz3127	FORWARD	57.54	GTATCGGTGGTTTTGATGCT	67	(ATT)6
	REVERSE	58.13	CCTCCTCTGTCGGTTTCC		
Brz3128	FORWARD	57.51	TATGTGCCTTTCACTTGCAG	90	(AAT)8
	REVERSE	57.02	TTGAAGACAAGCACCTTGAA		
Brz3129	FORWARD	57.89	CCTTTTCCATGAACACTAGCA	75	(TCA)8
	REVERSE	57.18	ACTGGACTTTGTCGACCTGT		
Brz3130	FORWARD	58.14	TATGACATCGCCCTTTGC	71	(TCA)4
	REVERSE	58.11	AAGGAGCAGATGGAAGCTG		
Brz3131	FORWARD	57.72	GGAAAAATATTAACCGGCAAG	89	(CAA)5
	REVERSE	55.58	CTTGCATGGAGCTACAATTT		
Brz3132	FORWARD	57.34	GGCTCACTTACATGAAACAGG	75	(TGA)3
	REVERSE	58.35	GGTGCTTGTTGCTCATTTTC		
Brz3133	FORWARD	59.13	TCGACGATAAGAAGGGGAAC	65	(TGA)5
	REVERSE	58.62	CCTTTGTCAGCATTGGGTAA		
Brz3134	FORWARD	57.89	TTCGCCAAATTTACAATTTTCT	80	(TTG)4
	REVERSE	57.63	GCCAAACGTTACTTCAACTCTT		
Brz3135	FORWARD	57.74	GCTTCACTCTCTGCAAGACC	78	(TCA)3
	REVERSE	57.88	TCAGTAGATGAGCACGGATATG		
Brz3136	FORWARD	55.53	CAAGACAAGCCTGAACACTT	94	(CTT)11
	REVERSE	58.74	CCAGCCGTGCAATACTATACA		
Brz3137	FORWARD	58.08	TGGTGCTAGCGTATCTCCTC	72	(TTC)4

	REVERSE	58.10	CAGCTACCAGTTTCATCAGTAGG		
Brz3138	FORWARD	57.33	CGTTAGACGAAAGGGACAAC	100	(TTC)3
	REVERSE	56.47	GGAACCACTTCTATTCTCTTTTCT		
Brz3139	FORWARD	57.80	TCTCCTCGCTTTAGGTCTTATG	91	(CTT)3
	REVERSE	57.87	CTTTGCCTGTCTCTGTCCAT		
Brz3140	FORWARD	58.57	ATCCTGCAGGTCGATCCTAT	97	(TTC)7
	REVERSE	58.50	TACCAACACTCCTGGTACGG		
Brz3141	FORWARD	58.80	CGCTCATCGTTACGACATTT	95	(CTT)4
	REVERSE	58.39	TGCCTGGAGGAGATTGAGTA		
Brz3142	FORWARD	55.14	CAACAAGAGAAGAGATATGAAGC	92	(AAG)4
	REVERSE	56.73	CTGTGTGTTGATCATCCTTTTT		
Brz3143	FORWARD	58.28	GGGACTACGGTGTGTGTCAT	100	(TAA)4
	REVERSE	56.73	GCCTCACCTGTTTAGCTAGG		
Brz3144	FORWARD	57.78	GACGAAGAGGATGAACAGGA	73	(GAA)5
	REVERSE	57.93	ACGTGCGAGCATTTTTATTC		
Brz3145	FORWARD	57.99	TTCCTCCCACCATTGAGTAA	59	(TCA)4
	REVERSE	58.14	GAATTGTGAAGATGGCGTTC		
Brz3146	FORWARD	57.98	TGTGGTGTTTGACCTGTTTG	93	(CTT)6
	REVERSE	56.88	GAGATTTGCCTCCCTGTAAA		
Brz3147	FORWARD	58.60	GCCACACATACTCCACTCGT	91	(GAA)6
	REVERSE	56.87	AAAGGATGATTTGCCTTACG		
Brz3148	FORWARD	56.65	GGCGGCTTCATTAATTTTAG	88	(AAT)8
	REVERSE	57.47	CAAGACACGGTTGTGAGCTA		
Brz3149	FORWARD	56.50	TGCACTCTTGGATGTTATACACT	98	(TCA)6
	REVERSE	57.41	GGTGGTGAAGCTCAAGACTC		
Brz3150	FORWARD	56.98	CAAGCCTCTATACGGGTACAA	77	(CTT)9
	REVERSE	58.03	TCTAAGCCCAAAACAGCAAC		
Brz3151	FORWARD	57.55	TGATATGAATCCTCCAGTGTTG	70	(GTT)4
	REVERSE	56.30	TTCCTGAAGAGCAATTAGCA		
Brz3152	FORWARD	57.94	AGAAACATGACCGGTATAGGG	83	(AAC)3
	REVERSE	57.40	GGAAATATATTGGCGGATGA		
Brz3153	FORWARD	58.17	TGTTCAGCATCCAAAGGTG	99	(GAA)4
	REVERSE	56.52	GAAGCGACACGTGAAGAATA		
Brz3154	FORWARD	54.64	CCATGTAAGTCTTCAATGACC	100	(ATT)3
	REVERSE	58.88	GCAACAGCTCCTGGAACATA		
Brz3155	FORWARD	57.70	CTCCCCATGTTAAGCTGGTA	98	(TTG)3
	REVERSE	57.11	ACAACCCATCTTTGCTCAAT		
Brz3156	FORWARD	57.88	CGCTTCTCCTCCTTCTTCTT	85	(TTC)6
	REVERSE	58.98	GGATCAACATCAAACGCAAC		
Brz3157	FORWARD	57.74	GGCCCAAGACAGAAGATACA	100	(TCA)4
	REVERSE	59.86	GCCACGGTTGGTAAGATTGT		

Brz3158	FORWARD	58.41	TGACATCCAGGGTTGTTAGG	73	(TTG)3
	REVERSE	58.34	ATATTCGCCGCACTGTTCT		
Brz3159	FORWARD	51.02	TTTACGAAGAGATTCTTACTATTT	80	(TAA)3
	REVERSE	57.81	TCCTCATTAGATTCTCGTGGTT		
Brz3160	FORWARD	57.17	TATATCGCGCAAGAGACAAA	84	(AAC)4
	REVERSE	58.22	GACCGCAGTGGTTAAGTGTC		
Brz3161	FORWARD	57.74	AAGTGCTCCAATAGCAGTAAGG	82	(TTC)4
	REVERSE	57.08	CAAAGATCCTTTTAGACCAAGG		
Brz3162	FORWARD	58.34	CATATTGCTCATGCAGAGGTC	80	(AAG)5
	REVERSE	57.87	AGGGACAATGCTTGACTCAG		
Brz3163	FORWARD	58.07	CATGTCGAATGTCCAGCATA	72	(TGA)3
	REVERSE	58.27	TTGTGGTCTCAGTCGTTCCT		
Brz3164	FORWARD	57.84	TCTGCTACCGACAACATTGA	93	(AAT)3
	REVERSE	57.08	TTTGAGCATGTTTCATCCAC		
Brz3165	FORWARD	57.17	AAAAATCATAGCCTGCCTTCT	83	(TAA)5
	REVERSE	56.15	TGAGAATGAATAACTCGAGCA		
Brz3166	FORWARD	57.68	TCAGATGCAATGGAGAATCA	72	(CAA)4
	REVERSE	58.49	CCCGGTTCAAGGAGGTAA		
Brz3167	FORWARD	58.04	CGGATATAAGCACAACGCTAA	82	(TCA)7
	REVERSE	57.08	GCTTTGTAAAACGAGGCAAT		
Brz3168	FORWARD	57.79	AGGAGGTACAAGCCAAAGGT	93	(AAG)5
	REVERSE	57.61	TTTTGTCTTCGCCTTCACTT		
Brz3169	FORWARD	57.58	TATCCAAGGTTTGGGCTATG	65	(AAT)5
	REVERSE	57.79	CTTCTATGGCCATGCATCTT		
Brz3170	FORWARD	57.60	TCACAGAAAGAGCACAACATCT	90	(TTG)4
	REVERSE	57.77	TTGGAACAAGAAGGTTCGTC		
Brz3171	FORWARD	58.03	AGCCTGAAAATGTTGCAAAG	82	(GAA)4
	REVERSE	58.04	TCACTGCACACAAAGCAGTT		
Brz3172	FORWARD	57.88	CAAAAGGTGAAGCCAAGGTA	100	(TGA)4
	REVERSE	56.09	TTTGGTAAAACATAAGAGGGAAT		
Brz3173	FORWARD	58.33	TGTTGTGCGACACCTTAGTG	86	(TGA)3
	REVERSE	58.28	TCGGCATGAGTACAAGTGTG		
Brz3174	FORWARD	58.86	CTGGTTTATCAGGGGACGAT	85	(GAA)3
	REVERSE	57.14	TGATAGGAGGTTAGCAAGTCG		
Brz3175	FORWARD	57.80	GAAAGGTGAAGCCAAGGTATC	93	(TGA)4
	REVERSE	56.00	TCATCATATAAGGGAATATCGTCT		
Brz3176	FORWARD	57.97	AACAAGACAACCATGGAGGA	72	(AAG)7
	REVERSE	57.71	GGAAGAGATTCAAACCGTGA		
Brz3177	FORWARD	57.88	CAAAAGGTGAAGCCAAGGTA	94	(TGA)4
	REVERSE	57.91	TCATCATATGAGGGAATATCGTC		
Brz3178	FORWARD	58.08	GCAGCATGCAACTAGGAGAT	92	(AAG)7

	REVERSE	57.55	GGATTTACCATTTGTTTCTTGG		
Brz3179	FORWARD	57.71	TGGATAAGATCAAAAGCATGG	69	(AAG)4
	REVERSE	58.23	TGGTCGTCATGTTAGGCTTT		
Brz3180	FORWARD	58.39	GCCTCTTCTTCCACGAATTT	69	(TTC)4
	REVERSE	57.86	TGCATTGGAGCAATAGTTGA		
Brz3181	FORWARD	57.28	CGAGCCTATTAATGAAGCAAA	68	(AAC)3
	REVERSE	59.04	CCAGTTTATGGCCCCTTCTA		
Brz3182	FORWARD	59.02	GGCCATGATCAGGTCAAAG	82	(GAA)3
	REVERSE	58.65	TGCCTTTCCTTCTACCCATT		
Brz3183	FORWARD	58.45	ATCTCGGGAAAGGTATTCCA	72	(TGA)5
	REVERSE	58.33	CGGAGGGAGAATAAAGGAGA		
Brz3184	FORWARD	57.61	TTTGGACACCAAACTCTTCC	83	(CTT)3
	REVERSE	58.17	CCAATGGGTGTGATTTTTGT		
Brz3185	FORWARD	58.09	AATTGGGTTCATAAGCACGA	69	(AGA)5
	REVERSE	57.30	TTTGTTATGTCACATGCAAGC		
Brz3186	FORWARD	58.22	ACACAGTTGCTTCCGATTGT	95	(AAT)5
	REVERSE	55.99	GCGATGCATTGCTAAAAA		
Brz3187	FORWARD	58.50	CAACTTTGTCTTGCCCAGAG	71	(AAT)5
	REVERSE	57.73	ATTGGCAAAAATCTCCTCCT		
Brz3188	FORWARD	53.72	AACTATGCCCATTTTTATATTTC	97	(ATT)5
	REVERSE	57.92	CGCACATACGAAGGAGAGAT		
Brz3189	FORWARD	50.87	AGTTAGAGTTATTATCTTTTCCAA	72	(TAA)4
	REVERSE	56.39	TCAAGATACAACTTTCCTCAGTCT		
Brz3190	FORWARD	57.47	CAACATGTGGTCGTTCACTC	94	(ATT)8
	REVERSE	57.94	ACTGATGCCTCATCCAATGT		
Brz3191	FORWARD	58.70	CGTTTTACGTACGGTCCTGA	89	(AAT)5
	REVERSE	51.76	TCTCAGAGAGCATAGTTTATTG		
Brz3192	FORWARD	56.95	TCGAATTGGATCATCATCTG	98	(GAA)5
	REVERSE	58.32	CATGTGTGATTGGATGTTGC		
Brz3193	FORWARD	58.30	GGCATAATTTTGGCATGGT	69	(CAA)4
	REVERSE	59.15	CCGTCGTGCATCAGAAATAC		
Brz3194	FORWARD	58.05	GTCATCATCGGCATCTTCA	71	(TGA)6
	REVERSE	58.13	GCTGCTTCTCTTGCTTCTTG		
Brz3195	FORWARD	59.06	TCGAGGTAGCCTCCAAGC	85	(ATT)5
	REVERSE	58.08	ACGGCAGCATATCACATTG		
Brz3196	FORWARD	56.90	AAACGAACTTCCCACTCAAG	81	(CTT)7
	REVERSE	57.97	CTGTAACCCTGCCCAATATG		
Brz3197	FORWARD	58.26	AAAGAGGGATTCCCAGAGC	61	(GAA)3
	REVERSE	57.45	GCAGGGATGAGAGTCAAAAG		
Brz3198	FORWARD	57.08	CATGCATGAAAAGTATGTGGA	82	(ATT)6
	REVERSE	57.57	TGAAGCAATATAAAGCCAAACA		

Brz3199	FORWARD	57.65	GGATGGGGATTGGACTTTAT	73	(TTC)4
	REVERSE	57.98	TAGCAACCAACACAAAGCAA		
Brz3200	FORWARD	58.03	AATCTCTGCAAGTGGAGTCG	71	(TTC)5
	REVERSE	57.70	ATCAAAAGGTATGCCAGCAC		
Brz3201	FORWARD	58.12	ACAAGGTCGTGCTGTCTAGC	71	(AAG)6
	REVERSE	57.22	ATCCTTTTCAGAGGGAGGTC		
Brz3202	FORWARD	57.83	AATCATCGTGATCGGGAATA	81	(TCA)4
	REVERSE	57.80	ACGTGCCTCATTTTGAACTC		
Brz3203	FORWARD	58.03	GTGTACGAGAAGGCTGAGGA	77	(TTC)3
	REVERSE	58.75	TCGCTCCACACTCACACAT		
Brz3204	FORWARD	57.85	AACACACTTTTTGGCGACTC	78	(GAA)4
	REVERSE	57.64	ATAGCTCATTTGCCATCACC		
Brz3205	FORWARD	57.84	GCTGCCATTTAATTTCAGGA	129	(ATT)7
	REVERSE	58.70	CGTTGCTACGGAACAACATT		
Brz3206	FORWARD	56.46	GGCCATTATTGTTGTTGTTG	118	(TTA)6
	REVERSE	50.83	ACTAACTTAATGGTAAGCAAAA		
Brz3207	FORWARD	50.05	TTTAAAAGAATAAAGTTAATACGA	76	(TAA)3
	REVERSE	53.11	TTCGTGTTCCACTAATCAAT		
Brz3208	FORWARD	58.26	AGGAGGAGAGGTGGAATCTG	74	(AAG)7
	REVERSE	58.99	CCGCCTTTTTCTCCTTCTC		
Brz3209	FORWARD	58.38	TGTAGGTGGTTTTGGTTTGG	90	(TTG)5
	REVERSE	58.46	GCAAAGTTAAGCTTGCGAGA		
Brz3210	FORWARD	50.40	ATAAACCTTTTATTTCTAGCTTT	98	(AAT)5
	REVERSE	57.63	CACATTAAAAAGGACTCTATGCAA		
Brz4011	FORWARD	57.34	TTTTTGATTTGCCCTGAACT	107	(TTTC)3
	REVERSE	58.02	AAAAGAAAACACCCCGAATC		
Brz4012	FORWARD	58.06	TCGTTGGAAAGACCCTGAT	101	(TTTC)3
	REVERSE	57.65	CCTCCCTACTCATCCAGTCA		
Brz4013	FORWARD	58.06	TTCATGCGCTCATTAAATAAGTT	94	(ATTT)3
	REVERSE	57.48	GGTAAGGTAAAAACAATGCAAGA		
Brz4014	FORWARD	57.27	CCTGACAGCTGCATGTGTAT	115	(ATTT)3
	REVERSE	53.37	AATCCCTGTACGTTAAAGAAA		
Brz4015	FORWARD	59.09	AACGAGAGCTCGTCAACCTT	102	(GTAC)3
	REVERSE	57.75	GCACGGTTTCACGTTCTATT		
Brz4016	FORWARD	54.90	AGGTATCGCAGTAGAAGCAG	100	(TAAA)3
	REVERSE	53.50	TTTTGACAAATTGAAATGCT		
Brz4017	FORWARD	57.12	TGGCTACCTGATCATTTGTG	96	(CATG)3
	REVERSE	57.60	GAAAGAAGTGTGCACAAGCA		
Brz4018	FORWARD	57.73	TCTCCAATTTCAGAATGTGATTT	119	(ATTT)3
	REVERSE	52.24	CAATTGTTTTTATTTCTACGC		
Brz4019	FORWARD	52.77	CTTAGAATATGTATTCGACGTAAA	89	(ATTT)3

	REVERSE	57.60	AGCCAATTAAATATTTGCAAGC		
Brz4020	FORWARD	57.62	AGTGACTAGTTTGACGGCTCA	84	(AAAG)4
	REVERSE	57.57	TCGTGGAATGCATTTTCATA		
Brz4021	FORWARD	57.35	GAGAGGAGTGACATGGAAGC	113	(CATG)4
	REVERSE	57.59	TCCCCTCCTCTCTTATCCAT		
Brz4022	FORWARD	57.46	CCCTTATTGGGATGCATAAA	111	(ATTT)4
	REVERSE	57.51	TGCATGATCTATACTGCGTTTT		
Brz4023	FORWARD	57.67	CAAATGCATCCATACAACTTACA	117	(TAAA)4
	REVERSE	58.73	CCATTTTACAGCCTCCATAGC		
Brz4024	FORWARD	57.67	AAATCGGAAACTTGTGATGC	99	(TAAA)4
	REVERSE	57.98	ATACTTCGCTGTTGCCACTC		
Brz4025	FORWARD	58.47	CCCAGAAATTTGACCGTGTA	102	(AAAG)4
	REVERSE	57.22	GGCCATTGACCAACTTTTAG		
Brz4026	FORWARD	58.09	AATAGGCCCTTATTGGTTGC	102	(TAAA)4
	REVERSE	57.97	AACTACCACCACACCACTTCA		
Brz4027	FORWARD	58.12	ATGCAGGCAATCAGATCAAT	117	(TGCA)4
	REVERSE	57.63	GTAGCGTGCAAAGTTCCTTC		
Brz4028	FORWARD	54.39	AACAAATGAGAAAGAAAGAAAAA	100	(CATG)4
	REVERSE	57.65	GCAGTGCTGTGATGGACA		
Brz4029	FORWARD	57.92	AAACATATTTCACCGGGACTC	114	(TGCA)4
	REVERSE	58.65	GTGTGCACCCAAAAAGAAAA		
Brz4030	FORWARD	58.60	ATTGATTGGTTGGGCTCTCT	109	(ATTT)4
	REVERSE	57.30	GGGAGTATTCCCATTTCACA		
Brz4031	FORWARD	58.43	ATGTGGAGGGCAGGAAAC	87	(TGCA)4
	REVERSE	57.05	CTCAGTCACAAGAGCACGAC		
Brz4032	FORWARD	57.03	GGTGGCGGAATATGTCTTAC	85	(ATTT)3
	REVERSE	57.58	AATGGTGTTTTGGACTTTGC		
Brz4033	FORWARD	58.19	ATCCAAACCCCATCGTCTAT	94	(CATG)4
	REVERSE	57.42	TTTTGATCGCTTTGTTGTTG		
Brz4034	FORWARD	57.82	TCCCAAACTCCAAGCAATAG	92	(TAAA)3
	REVERSE	57.77	AGTCAACTCAACTGGCAACC		
Brz4035	FORWARD	58.29	TCGATGTAACGAATGCACAG	111	(CATG)3
	REVERSE	59.35	CGCATGCATGTTTAGCTACC		
Brz4036	FORWARD	56.81	TTTGAACTGTTGACATGTGCT	80	(ATTT)3
	REVERSE	57.30	GTTCATAAATGGGTGGGAGA		
Brz4037	FORWARD	57.46	TGCGAATTTACCTCGTTTTT	83	(GTAC)4
	REVERSE	58.38	CGAAACTACAAGCACATACGG		
Brz4038	FORWARD	58.44	ATAGGAGCTTGGGCAGTAGG	118	(AAAG)4
	REVERSE	59.40	CGTCAAACTGTGGGCTCTC		
Brz4039	FORWARD	58.05	CCAGATCCACAGTGACCTTC	83	(AAAG)3
	REVERSE	58.00	TTCCTTTGCTTTCTTTGTGG		

Brz4040	FORWARD	56.49	TTTGAAAATTTGGGATGTGA	115	(TTTC)3
	REVERSE	57.42	AGCAATCCGTCCAGAAGTTA		
Brz4041	FORWARD	58.17	ATGACCTATTTGGACGAGCA	82	(CATG)4
	REVERSE	57.95	GTTTGAGCAGCCAATGTCTT		
Brz4042	FORWARD	58.86	CATGCCAGGCTGTCTTTTT	94	(AAAG)4
	REVERSE	57.17	TCACGCTTAAGTTTCAACGA		
Brz4043	FORWARD	58.92	TGGTTTTATGCCTCAGCTTG	111	(ATTT)3
	REVERSE	56.04	CCATGATTCAAGAGATTAAACAA		
Brz4044	FORWARD	57.69	GGCGCTGCTATCTTGTAAAC	67	(CATG)3
	REVERSE	58.39	CATAGCACGCAATGCATAAA		
Brz4045	FORWARD	57.88	CGCAACACTTTTTGAGGAAT	78	(TTTC)3
	REVERSE	58.20	TCTTGACAAATGCCCTGTTT		
Brz4046	FORWARD	57.44	GAGAACATGCACCATGATTG	84	(TAAA)3
	REVERSE	58.22	CCTTAACGTCCTGAGTGTGG		
Brz4047	FORWARD	57.74	ATCAGGCATCAAACGAAGAC	113	(ACGA)3
	REVERSE	58.28	TTGATCCCCATCAAAATACG		
Brz4048	FORWARD	57.52	TGTGAGCTTATTTAAATTTGATGG	79	(TGCA)3
	REVERSE	58.19	CAGCTGTCCAGCAGAAGAAT		
Brz4049	FORWARD	58.39	ATGGCATGTCTCGCTAAAAA	87	(TTTC)4
	REVERSE	57.14	CCGGTGCTGAAATGACTT		
Brz4050	FORWARD	57.95	TCGGGCAAAGTTAGTTTTTG	91	(AAAG)4
	REVERSE	58.34	AAACACTTCCAGCACTTCCA		

Additional file 2

Descriptive statistics of *B. ruziziensis* microsatellite markers

Additional file 2

Descriptive statistics of *B. ruziziensis* microsatellite markers

	Descriptive Statistics									Transferability				
Marker	**Allele Ferquency**	**Genotype No**	**Sample Size**	**No. of obs.**	**Allele No**	**Observed size ranges**	**He**	**Ho**	**PIC**	***Briz1***	***Briz2***	***Briz3***	***Dec***	***Hum***
Brz0001	0.6818	3	11	11	2	150-152	0.406	0.273	0.340	+	-	+	+	-
Brz0002	0.9091		11	11	2	145-147	0.150	0.000	0.152	+	-	+	+	-
Brz0004	0.3333	6	11	9	6	113-155	0.703	0.222	0.745	+	+	+	+	+
Brz0007	0.4091	8	11	11	6	153-165	0.688	0.364	0.703	+	+	+	+	+
Brz0008	0.3500	7	11	10	6	151-165	0.718	0.900	0.711	+	+	+	+	-
Brz0009	0.5000	5	11	10	4	137-143	0.540	0.300	0.501	+	+	+	+	+
Brz0010	0.9444	2	11	9	2	157-159	0.099	0.111	0.099	+	+	+	+	+
Brz0011	0.6364	4	11	11	3	131-135	0.490	0.545	0.444	+	+	+	+	+
Brz0012	0.5000	6	11	11	5	150-162	0.613	0.273	0.613	+	+	+	+	+
Brz0013	0.4545	5	11	11	3	148-152	0.586	0.364	0.551	+	-	+	+	-
Brz0014	0.8636	2	11	11	2	150-152	0.226	0.273	0.208	+	+	+	+	+
Brz0015	0.2727	6	11	11	6	144-156	0.764	0.636	0.783	+	+	+	+	-
Brz0017	0.1818	10	11	11	11	134-160	0.840	1000	0.864	+	+	+	+	-
Brz0019	0.9500	2	11	10	2	148-150	0.090	0.100	0.090	+	+	+	+	+
Brz0021	0.2500	8	11	10	9	151-183	0.770	0.400	0.816	+	+	+	+	-

Brz0023	0.5000	4	11	4	3	125-145	0.498	0.500	0.511	+	+	+	+	+
Brz0024	0.3182	8	11	11	5	145-155	0.711	0.545	0.715	+	-	+	+	-
Brz0026	0.8500	3	11	10	3	157-163	0.253	0.300	0.247	-	-	-	+	+
Brz0027	0.7500	4	11	10	4	141-149	0.393	0.500	0.379	+	+	+	+	+
Brz0028	0.4375	5	11	8	4	153-161	0.589	0.375	0.582	+	+	+	+	+
Brz0029	0.3750	7	11	8	7	133-155	0.706	0.625	0.735	+	+	+	+	+
Brz0030	0.3500	7	11	10	5	140-152	0.667	0.600	0.658	+	+	+	+	-
Brz0031	0.2222	9	11	9	7	144-156	0.770	0.667	0.804	+	+	+	+	+
Brz0032	0.7500	4	11	8	3	151-169	0.363	0.250	0.354	+	-	+	+	-
Brz0033	0.6818	5	11	11	4	136-154	0.447	0.273	0.427	+	+	+	+	+
Brz0034	0.6364	6	11	11	4	157-169	0.508	0.364	0.496	+	+	+	+	-
Brz0035	0.3571	5	11	7	5	146-154	0.670	0.857	0.666	+	+	+	+	+
Brz0036	0.9000	2	11	10	2	137-139	0.162	0.000	0.164	+	-	+	+	-
Brz0037	0.5455	7	11	11	4	143-149	0.591	0.545	0.574	+	+	+	+	+
Brz0038	0.2727	9	11	11	6	140-154	0.772	0.909	0.778	+	+	+	+	-
Brz0039	0.5000	5	11	10	4	139-151	0.605	0.700	0.573	+	+	+	+	-
Brz0041	0.8636	3	11	11	2	141-149	0.218	0.091	0.208	+	-	+	-	-
Brz0042	0.6667	4	11	6	3	143-147	0.430	0.333	0.424	+	+	+	+	-
Brz0043	0.3750	4	11	4	4	131-197	0.564	0.250	0.667	+	+	+	+	+
Brz0045	0.5455	6	11	11	7	142-154	0.633	0.727	0.635	+	-	+	+	-
Brz0047	0.3333	9	11	9	7	150-170	0.731	0.556	0.762	-	-	-	-	+
Brz0048	0.2273	10	11	11	9	139-161	0.813	0.818	0.839	+	-	+	+	-

Brz0049	0.9444	2	11	9	2	148-162	0.099	0.111	0.099	+	-	-	+	-
Brz0050	0.6818	5	11	11	3	151-155	0.448	0.364	0.419	+	+	+	+	+
Brz0051	0.5000	5	11	10	4	140-146	0.619	0.800	0.587	+	-	+	+	-
Brz0052	0.3333	5	11	6	5	149-157	0.650	0.500	0.692	+	-	+	+	-
Brz0055	0.5909	6	11	11	5	148-160	0.557	0.545	0.541	-	-	-	-	-
Brz0056	0.6818	5	11	11	4	141-149	0.458	0.091	0.467	+	-	+	+	-
Brz0058	0.5000	4	11	9	3	131-135	0.585	0.778	0.535	+	-	+	+	-
Brz0059	0.7273	4	11	11	4	133-143	0.422	0.455	0.411	+	+	+	+	+
Brz0060	0.4000	7	11	10	5	143-157	0.676	0.500	0.681	+	+	+	+	-
Brz0061	0.7273	2	11	11	2	147-149	0.385	0.545	0.318	+	-	+	+	-
Brz0062	0.5500	5	11	10	3	148-152	0.564	0.600	0.528	+	+	+	+	-
Brz0063	0.7727	4	11	11	4	131-145	0.369	0.455	0.362	+	+	+	+	+
Brz0064	0.3636	6	11	11	6	143-153	0.724	1000	0.709	+	+	+	+	+
Brz0065	0.2000	10	11	10	12	130-166	0.829	0.700	0.875	-	+	-	-	-
Brz0066	0.6500	5	11	10	4	140-148	0.510	0.600	0.498	+	+	+	+	-
Brz0067	0.5000	6	11	10	5	149-161	0.631	0.400	0.642	+	+	+	+	+
Brz0069	0.6364	5	11	11	4	147-155	0.520	0.636	0.496	+	+	+	+	-
Brz0070	0.5000	5	11	11	4	138-150	0.618	0.727	0.586	+	+	+	+	+
Brz0071	0.4444	6	11	9	6	135-151	0.673	0.889	0.661	+	+	+	+	+
Brz0072	0.5556	4	11	9	3	129-135	0.528	0.444	0.489	+	-	+	+	-
Brz0073	0.5556	6	11	9	6	110-158	0.588	0.333	0.610	+	+	+	+	-
Brz0075	0.2273	11	11	11	8	129-153	0.809	0.727	0.839	+	+	+	-	-

Brz0076	0.3333	6	11	9	5	145-167	0.686	0.333	0.712	+	+	+	+	+
Brz0077	0.4545	4	11	11	3	133-137	0.541	0.364	0.486	+	-	+	+	-
Brz0078	0.5000	5	11	10	3	152-156	0.546	0.500	0.492	+	+	+	+	+
Brz0079	0.4545	5	11	11	3	149-153	0.603	0.727	0.551	+	-	-	+	-
Brz0080	0.6667	2	11	9	2	146-148	0.395	0.000	0.346	+	-	+	+	-
Brz0081	0.7273	4	11	11	4	222-244	0.426	0.545	0.411	+	-	+	+	-
Brz0082	0.3889	6	11	9	4	251-261	0.615	0.444	0.602	-	+	+	-	-
Brz0083	0.2222	8	11	9	6	233-249	0.765	0.778	0.788	+	+	+	+	-
Brz0085	0.4167	4	11	6	5	240-248	0.603	0.333	0.644	+	+	+	+	+
Brz0087	0.3000	5	11	5	6	239-273	0.716	1000	0.748	+	+	+	+	+
Brz0089	0.3571	7	11	7	7	224-248	0.710	0.571	0.759	-	-	-	-	-
Brz0090	0.5455	3	11	11	2	257-261	0.466	0.364	0.373	+	+	+	+	-
Brz0092	0.3182	8	11	11	5	244-254	0.730	0.545	0.742	+	+	+	+	+
Brz0094	0.8000	4	11	10	3	239-253	0.311	0.200	0.303	+	+	+	+	+
Brz0096	0.5000	4	11	10	3	241-247	0.541	0.400	0.492	+	-	+	+	-
Brz0097	0.3125	7	11	8	5	234-242	0.705	0.375	0.747	+	-	+	+	-
Brz0099	0.4545	6	11	11	3	236-240	0.602	0.455	0.567	+	-	-	+	-
Brz0100	0.5000	3	11	4	3	242-252	0.470	0.250	0.511	+	+	+	+	+
Brz0101	0.3750	5	11	8	4	248-254	0.630	0.375	0.636	+	+	+	+	-
Brz0102	0.8182	3	11	11	3	245-251	0.298	0.364	0.282	+	+	+	+	-
Brz0103	0.9500	2	11	10	2	248-250	0.090	0.100	0.090	-	-	+	+	+
Brz0104	0.9444	2	11	9	2	242-244	0.099	0.111	0.099	+	+	+	+	-

Brz0105	0.5000	2	11	2	2	244-246	0.250	0.000	0.375	+	+	+	+	+
Brz0107	0.2222	9	11	9	6	227-247	0.747	0.444	0.788	+	-	+	+	-
Brz0108	0.6667	4	11	9	3	251-255	0.450	0.222	0.438	+	+	+	+	-
Brz0109	0.4545	8	11	11	4	251-267	0.643	0.364	0.641	-	-	+	+	-
Brz0110	0.7500	4	11	10	4	243-249	0.392	0.400	0.389	+	-	+	+	-
Brz0111	0.5000	4	11	6	4	248-256	0.568	0.333	0.599	+	+	+	+	-
Brz0112	0.5000	5	11	10	4	236-244	0.595	0.400	0.581	+	+	+	+	-
Brz0113	0.7500	3	11	10	2	243-245	0.352	0.300	0.305	+	-	-	+	-
Brz0114	0.4375	6	11	8	6	245-275	0.672	0.625	0.691	+	+	+	+	-
Brz0115	0.5909	5	11	11	4	230-258	0.507	0.273	0.471	+	+	+	+	+
Brz0116	0.2500	10	11	10	9	223-271	0.803	0.900	0.827	+	-	+	+	-
Brz0117	0.5000	5	11	7	4	256-262	0.597	0.571	0.599	+	+	+	+	+
Brz0118	0.2273	10	11	11	10	237-263	0.812	0.636	0.849	+	+	+	+	+
Brz0119	0.8333	3	11	9	2	244-246	0.253	0.111	0.239	+	+	+	+	+
Brz0120	0.3750	7	11	8	5	242-250	0.685	0.500	0.711	+	+	+	+	-
Brz0121	0.6000	4	11	10	3	245-249	0.477	0.300	0.424	+	-	+	+	-
Brz0122	0.5000	5	11	11	3	250-256	0.583	0.545	0.542	+	+	+	+	+
Brz0123	0.3750	6	11	8	4	243-253	0.658	0.625	0.658	+	+	+	+	-
Brz0127	0.3750	4	11	4	3	243-249	0.544	0.500	0.582	+	+	+	+	-
Brz0128	0.5714	4	11	7	4	253-261	0.525	0.000	0.570	+	+	+	+	-
Brz0129	0.8889	2	11	9	2	256-258	0.176	0.000	0.178	+	+	+	+	-
Brz0130	0.1818	9	11	11	9	242-266	0.820	0.636	0.858	+	+	+	+	+

Brz0131	0.3571	6	11	7	6	254-264	0.712	0.714	0.744	+	+	+	+	-
Brz0132	0.4444	7	11	9	6	246-264	0.688	0.778	0.695	+	+	+	+	-
Brz0134	0.7273	4	11	11	3	247-251	0.395	0.364	0.360	+	+	+	+	+
Brz0136	0.7500	3	11	6	3	240-246	0.361	0.333	0.363	+	+	+	+	+
Brz0138	0.4091	9	11	11	6	245-259	0.689	0.636	0.687	+	-	+	+	-
Brz0139	0.5909	6	11	11	6	236-246	0.560	0.273	0.569	+	-	+	+	-
Brz0140	0.4167	5	11	6	7	236-256	0.673	0.500	0.739	+	+	+	+	-
Brz0142	0.1875	8	11	8	10	241-287	0.823	0.875	0.871	+	+	+	+	-
Brz0143	0.6364	3	11	11	3	204-238	0.481	0.000	0.473	+	+	+	+	-
Brz0144	0.5500	5	11	10	6	245-261	0.590	0.300	0.604	+	+	+	+	-
Brz0145	0.6111	4	11	9	3	255-259	0.483	0.111	0.468	+	+	+	+	+
Brz0147	0.2000	9	11	10	10	240-288	0.825	0.700	0.868	+	-	+	+	-
Brz0148	0.2727	10	11	11	8	248-274	0.787	0.909	0.800	+	-	-	+	-
Brz0149	0.3750	4	11	8	3	231-251	0.588	0.250	0.582	+	+	+	+	+
Brz0150	0.4000	6	11	10	5	240-252	0.644	0.400	0.643	+	-	-	+	-
Brz0151	0.3750	6	11	8	4	243-255	0.643	0.375	0.658	+	+	+	+	+
Brz0152	0.2778	8	11	9	6	228-248	0.748	0.667	0.774	+	+	+	+	+
Brz0153	0.5909	7	11	11	6	247-257	0.568	0.455	0.569	+	+	+	+	-
Brz0156	0.3182	10	11	11	7	239-265	0.752	0.636	0.770	+	+	+	+	+
Brz0157	0.5000	7	11	10	5	234-246	0.636	0.600	0.634	+	+	+	+	+
Brz0158	0.8500	3	11	10	3	244-248	0.243	0.100	0.247	-	-	-	+	-
Brz0160	0.3000	5	11	5	5	240-252	0.680	0.800	0.720	+	+	+	+	+

Brz0161	0.5000	8	11	11	5	285-293	0.639	0.455	0.645	+	-	+	+	-
Brz0162	0.7500	4	11	8	3	288-292	0.363	0.250	0.354	+	+	+	+	+
Brz0163	0.8571	3	11	7	3	301-305	0.238	0.286	0.240	+	+	+	+	-
Brz0164	0.8182	3	11	11	2	296-298	0.278	0.182	0.253	+	+	+	+	+
Brz0165	0.5833	3	11	6	2	294-296	0.443	0.500	0.368	+	-	+	+	+
Brz0166	0.7500	2	11	4	3	294-298	0.330	0.250	0.371	+	-	+	+	-
Brz0167	0.9091	3	11	11	3	289-299	0.162	0.182	0.163	+	+	+	+	+
Brz0168	0.7857	3	11	7	3	304-310	0.315	0.143	0.325	+	+	+	+	+
Brz0169	0.6250	5	11	8	5	295-307	0.528	0.500	0.539	-	-	-	+	+
Brz0170	0.4375	6	11	8	4	282-298	0.596	0.500	0.582	+	+	+	+	-
Brz0171	0.2500	6	11	8	7	312-348	0.725	0.250	0.787	+	+	+	+	-
Brz0173	0.5556	2	11	9	2	290-292	0.439	0.000	0.372	+	-	+	+	-
Brz0174	0.5000	4	11	10	3	313-317	0.583	0.800	0.527	+	+	+	+	-
Brz0175	0.6364	5	11	11	4	286-296	0.515	0.364	0.511	+	-	-	+	+
Brz0177	0.2500	7	11	8	6	291-319	0.725	0.125	0.795	+	-	+	+	-
Brz0178	0.3333	3	11	3	3	304-320	0.444	0.000	0.593	+	+	+	+	-
Brz0179	0.6667	4	11	9	3	293-297	0.462	0.333	0.449	+	+	+	+	+
Brz0180	0.2727	8	11	11	7	285-305	0.751	0.091	0.800	+	+	+	+	+
Brz0181	0.6250	5	11	8	3	293-299	0.486	0.375	0.468	+	+	+	+	+
Brz0182	0.1818	10	11	11	10	252-328	0.823	0.545	0.868	+	+	+	+	+
Brz0183	0.6364	3	11	11	2	292-296	0.436	0.364	0.356	+	+	+	-	-
Brz0184	0.9091	2	11	11	2	300-302	0.158	0.182	0.152	+	+	+	+	+

Brz0185	0.7222	3	11	9	2	281-285	0.374	0.333	0.321	-	-	-	-	-
Brz0186	0.4000	3	11	5	3	291-297	0.512	0.000	0.563	+	-	+	+	-
Brz0190	0.7727	4	11	11	3	285-289	0.354	0.273	0.344	+	+	+	+	+
Brz0194	0.6500	4	11	10	3	305-307	0.478	0.500	0.442	+	+	+	+	-
Brz0195	0.3889	6	11	9	5	283-291	0.670	0.333	0.693	+	+	+	+	+
Brz0196	0.9091	2	11	11	2	252-298	0.150	0.000	0.152	+	-	+	+	-
Brz0197	0.9375	2	11	8	2	290-292	0.110	0.125	0.110	+	+	+	+	-
Brz0198	0.5909	4	11	11	3	297-301	0.512	0.455	0.463	+	-	+	+	-
Brz0199	0.3333	3	11	3	4	295-303	0.522	0.333	0.671	-	-	-	-	-
Brz0201	0.5455	5	11	11	4	204-292	0.553	0.364	0.522	+	+	+	+	+
Brz0202	0.5000	4	11	6	5	309-329	0.580	0.333	0.622	+	+	+	+	-
Brz0203	0.5455	6	11	11	5	290-300	0.583	0.545	0.562	+	+	+	+	+
Brz0204	0.5714	5	11	7	4	278-290	0.536	0.571	0.520	+	-	+	+	+
Brz0205	0.7000	3	11	10	2	291-293	0.397	0.400	0.332	+	+	+	+	+
Brz0206	0.2500	9	11	10	7	278-302	0.757	0.600	0.783	+	-	+	+	-
Brz0209	0.5833	4	11	6	4	292-300	0.522	0.333	0.552	-	-	-	-	+
Brz0211	0.5556	6	11	9	6	290-300	0.606	0.667	0.610	-	-	-	-	+
Brz0212	0.3333	8	11	9	5	290-312	0.703	0.556	0.721	+	+	+	+	+
Brz0213	0.4286	6	11	7	4	292-310	0.633	0.714	0.626	+	+	+	+	+
Brz0214	0.5000	3	11	10	3	307-315	0.522	0.000	0.492	+	+	+	+	+
Brz0215	0.7727	3	11	11	3	252-298	0.338	0.091	0.326	+	+	+	+	-
Brz0216	0.5000	3	11	4	3	282-298	0.469	0.000	0.555	+	+	+	+	-

Brz0218	0.9000	2	11	10	2	295-297	0.172	0.200	0.164	+	+	+	+	+
Brz0219	0.2727	8	11	11	6	294-304	0.768	0.818	0.778	+	-	-	+	-
Brz0220	0.5625	3	11	8	2	297-299	0.438	0.125	0.371	+	+	+	+	-
Brz0221	0.8750	2	11	8	2	283-291	0.191	0.000	0.195	+	+	-	+	-
Brz0222	0.8889	2	11	9	2	346-348	0.187	0.222	0.178	+	-	+	+	-
Brz0223	0.8000	2	11	10	2	276-288	0.307	0.400	0.269	+	+	+	+	-
Brz0224	0.8889	2	11	9	2	297-299	0.176	0.000	0.178	-	-	-	-	+
Brz0225	0.5000	2	11	2	2	316-320	0.250	0.000	0.375	-	-	-	-	-
Brz0228	0.5625	3	11	8	3	305-309	0.479	0.125	0.447	+	+	+	+	-
Brz0231	0.9000	3	11	10	3	294-298	0.176	0.200	0.177	+	+	+	+	-
Brz0232	0.5000	6	11	11	4	294-300	0.584	0.727	0.533	+	-	-	+	+
Brz0233	0.5625	4	11	8	3	296-300	0.520	0.375	0.496	+	+	+	+	+
Brz0234	0.7857	3	11	7	3	299-305	0.315	0.143	0.325	-	-	-	+	-
Brz0235	0.5000	8	11	11	6	284-300	0.659	0.636	0.664	+	+	+	+	+
Brz0238	0.5000	6	11	11	4	296-302	0.598	0.364	0.580	+	+	+	+	+
Brz0239	0.4000	7	11	10	4	301-307	0.667	0.500	0.665	+	+	+	+	+
Brz3001	0.4091	7	11	11	5	143-155	0.635	0.455	0.620	+	-	+	+	-
Brz3002	0.4091	7	11	11	4	140-152	0.681	0.727	0.666	+	+	+	+	+
Brz3003	0.3571	6	11	7	6	147-177	0.684	0.571	0.719	+	+	+	+	-
Brz3004	0.9091	3	11	11	3	138-153	0.162	0.182	0.163	+	+	+	+	+
Brz3006	0.8636	2	11	11	2	138-144	0.226	0.273	0.208	+	+	+	+	+
Brz3009	0.6000	5	11	10	4	122-140	0.518	0.200	0.509	+	+	+	+	+

Brz3010	0.7143	2	11	7	2	140-143	0.389	0.571	0.325	+	+	+	+	+
Brz3018	0.7857	3	11	7	3	237-246	0.335	0.429	0.325	+	+	+	+	-
Brz4001	0.8500	3	11	10	2	132-140	0.234	0.100	0.222	+	+	+	+	-
Brz4002	0.8636	3	11	11	3	144-160	0.234	0.273	0.228	+	+	+	+	-
Brz4003	0.6667	2	11	3	2	130-134	0.296	0.000	0.346	-	-	+	+	-
Brz4004	0.7500	4	11	10	3	145-157	0.379	0.300	0.368	+	+	+	+	-
Brz4009	0.3333	9	11	9	5	124-148	0.699	0.667	0.704	+	+	+	+	-

Briz1= B. brizantha cv. Marandu
Briz2= B. brizantha cv. Piatã
Briz3= B. brizantha cv. Xaraés
Dec= B. decumbens cv. Basilisk
Hum= B. humidicola cv. Tupi

IX. CAPíTULO 2

De novo genome assembly of ruzigrass (*Brachiaria ruziziensis*): a genomic view of a species belonging to the most planted forage genus in the tropics

CAPÍTULO 2

De novo genome assembly of ruzigrass (*Brachiaria ruziziensis*): a genomic view of a species belonging to the most planted forage genus in the tropics

Abstract

Only a few *Brachiaria* species are responsible for millions of hectares of pasture used as green feed and hay in the tropics. The area cropped with *Brachiaria* in Brazil alone exceeds 85 M ha. *Brachiaria ruziziensis* is one of the most important cultivated forage species, especially for integrated non-tillage cropping systems. Currently little is known about its genome. In contrast to other Polaes, the development of the first genomic tools is in progress and can provide support to *B. ruziziensis* breeding programs. Next-generation sequencing (NGS) technologies have been used for a *de novo* partial assembly and analysis of the *B. ruziziensis* genome based on paired- end Illumina data sequence reads. Sequence assembly was conducted for a database of 20,211,010,488 bp. This sequence data corresponds to ~33x coverage of the ruzigrass genome. The *de novo* assembly procedures culminated on a draft comprising ~218 Mbp, which corresponds to about 35% of the estimated *B. ruziziensis* genome size. A nearly non-redundant high quality reference gene set of the *B. ruziziensis* genome was obtained, which contain 22,554 sequences. A total of 17,245 gene orthologs were identified between the *B. ruziziensis* gene set and three grass species (sorghum, maize and switch grass). The estimate of the protein-coding genes indicated between 42,876 and 49,381 genes in the ruzigrass genome. A total of 430,846 di, tri- and tetra-nucleotide simple sequence repeats (SSRs) was identified. A set of 18,162 perfect SSRs was selected for use in genetic analysis and breeding of ruzigrass. The *B. ruziziensis* genome seems to have a smaller transposable element content than the rice genome. Millions of DNA sequence reads obtained in a single run of NGS equipment provided enough data to initiate the genomic analysis of ruzigrass.

Introduction

The genus *Brachiaria* (tribe Paniceae; subfamily Panicoideae; family Poaceae) contains about 100 species from tropical and subtropical regions of Africa, South America and Australia. Around seven perennial species of the *Brachiaria* genus have historically been used as forage plants in tropical America, Asia, the South Pacific and Australia [1]. It is probable that the land used with cultivated forage crops in the tropics extends for hundreds of millions of hectares. Cultivated pastures in Brazil, for instance, cover around 100 million hectares Segundo IBGE em 2011. It is estimated that four *Brachiaria* species (*B. decumbens*, *B. humidicola*, *B. brizantha,* and *B. ruziziensis*) cover 85% of the cultivated pastures in Brazil alone [2] .

One of these species, the sexual diploid *Brachiaria ruziziensis* (2n=2x=18), was introduced in Brazil in the 1960s [1, 3]. Its use was initially limited by its poor adaptation to low-fertility soils. Commercial cultivars of other species, first developed from introductions of wild germplasm, were successfully established and widely planted in Latin America, mainly due to their more favorable traits. These included the apomictic polyploids *B. decumbens*, *B. humidicola*, and *B. brizantha*. At least until the 1990s, only five accessions of four *Brachiaria* species had been used as sources for selection of a very limited number of 20 cultivars [3]. As a consequence, the genetic basis of *Brachiaria* pastures has been, from early introductions, extremely narrow. Additionally, the extensive use of apomictic clones of these polyploid species, in areas covering tens of millions of hectares, represents a high risk of genetic vulnerability in forage production.

New brachiaria cultivars with a broad genetic base must be developed and adopted for forage pasture diversification in the tropics. The risk of genetic vulnerability in brachiaria pastures could be reduced by an increased use of genetic diversity kept in germplasm banks. However, the fact that polyploid brachiaria species typically present apomictic reproduction is a limiting factor in their breeding programs. These programs would benefit from genetic recombination based on sexual crosses for the selection of superior genotypes. Ruzigrass could have a major role in this process since sexual crosses are easily performed in this species and can benefit the breeding program. Also, after chromosome duplication, tetraploid ruzigrass plants can be crossed with other brachiaria species, allowing the inter-specific introgression of genes for the generation

of brachiaria hybrids.

The demand for *B. ruziziensis* seeds in Brazil is increasing due to its common use in integrated forest-crop-livestock production systems. *B. ruziziensis* has good forage quality, and grows fast in the beginning of the rainy season. It is well adapted to overseeding, has smaller herbicide demands for drying prior to the establishment of the next crop, and a small tussock architecture [4]. It can be used for animal feeding, as green pasture or hay, and also as soil coverage in no-till farming systems. Flowering occurs once a year, so seed production is uniform. This decreases seed production costs and increases seed quality.

The relatively small size of the *Brachiaria ruziziensis* genome enables its investigation and analysis, and the development of genomic tools to aid in breeding programs for this species. While tetraploid brachiaria species (*B. decumbens* and *B. brizantha*) have larger and more complex genomes (> 1,600 Mbp), the estimated genome size of *B. ruziziensis* is ~600 Mbp [5], similar in size to model grass species such as rice (430 Mbp) and sorghum (700 Mbp). Genomic tools such as molecular markers would support ruzigrass breeding programs and stimulate a more dynamic development of new cultivars. However, little is known about the genomic features of this species, such as its number of genes and retro-elements content. Linkage or QTL maps are not available, and the collinearity of its genome with model species is not known. The first set of microsatellite markers for ruzigrass, for instance, has just recently been published by our group [6].

Next-generation sequencing (NGS) technologies have reduced sequencing costs and increased the generation of new sequence data. This is benefiting species for which little or no genomic information had been available until very recently [7]. They have been successfully applied to generate *de novo* assemblies for species with no prior availability of a reference genome. Some of these so-called “orphan” species, i.e., plant species with no or little genetic and/or genomic knowledge, now face a deluge of genomic sequence data, as in the case of pigeonpea [8], diploid cotton *Gossypium raimondii* [9], chickpea [10], and the rubber tree [11]. Genome and transcript sequence data will aid in the comprehension of biological phenomena implicated in crop breeding, such as heterosis and epigenetics [7]. The availability of genomic tools such as molecular markers (SSRs and SNPs, for instance) will impact breeding programs of these crops, allowing the construction of linkage maps, and facilitating trait mapping

and marker-assisted breeding.

NGS technologies were initially targeted at genome resequencing, mostly due to difficulties in assembling the large numbers of short sequence reads generated by some of these systems. *De novo* genome assembly is a computationally complex task, regardless of the sequencing technology used, falling into a category of mathematical problems for which no efficient solution is known [12, 13]. In addition to the massive number of reads and their short size, one of the main problems faced by assembly algorithms is the presence of repetitive elements, especially when these are longer than the length of a read [13].

Some recent papers describing the *de novo* assembly of plant genomes have either used NGS technologies alone, or in addition to data generated by Sanger sequencing (for BAC end sequencing, for instance). Examples of the former class of papers include the woodland strawberry *Fragaria vesca* [14], bread wheat [15], diploid cotton *Gossypium raimondii* [9], and the rubber tree *Hevea brasiliensis* [11]. The use of bacterial artificial chromosome (BAC) end sequences anchored to a marker-rich linkage map improved the *de novo* assembly of the chickpea genome, for which short-read sequence data was generated by Illumina sequencing [16]. This approach is similar to that used for the draft assembly of the *Theobroma cacao* genome [17]. These assembled genomes contained different proportions of the total genome sequences for each species. For the woodland strawberry, 87.5% (209.8 Mb of an estimated genome size of 240 Mb) were represented in scaffolds [13]. For *Gossypium raimondii*, this estimate was of 88.1 % [8]. Finally, for *Hevea brasiliensis*, this proportion reached only 52% [10].

Our objectives in this paper were the sequencing and *de novo* draft assembly of the *Brachiaria ruziziensis* genome. With this draft assembly, we expect to describe genomic regions which will be useful for the development of molecular and genetic tools that will assist in breeding programs for this species. These include microsatellite markers and SNPs, for instance. We will also gain an initial knowledge about the functional fraction of the ruzigrass genome, and about the complexity of its composition, describing repetitive and mobile elements, for example. Finally, we evaluate the potential utility of the data generated in this study in a further initiative for the complete genome assembly of this species.

Material and Methods

Plant material and genome sequencing - A self-pollinated *Brachiaria ruzizensis* plant (FSS-1 clone) was selected for genome sequencing since its expected increased homozygosity would facilitate the assembly process. The plant is maintained at Embrapa Gado de Leite, in Juiz de Fora (MG), Brazil. DNA was extracted from fresh young leaves using a standard CTAB protocol [18], with modifications as described [19]. The genomic library was prepared for sequencing according to manufacturer's instructions (www.illumina.com). In short, DNA was fragmented by nebulization and fragment 3´ends were ligated with A bases. DNA adaptors with a single T-base 3′-end overhang were ligated to the above products. Ligation products were run on 1% agarose gels and fragments of ~200 bp insert size were purified from the gel. Sequencing was performed from the genomic DNA fragment library, amplified by cluster generation by bridge PCR, allowing for the massive parallel paired end sequencing by synthesis using 3 channels of an Illumina GAII sequencer.

De novo genome assembly - The *B. ruziziensis* DNA sequence database was initially BLASTed against a database of chloroplast, mitochondrial and contaminant DNA (fungi, bacteria and virus) to verify the presence of non-nuclear and/or exogenous DNA. Also, every sequence above 700X coverage was inspected by BLAST checking in order to identify highly repetitive regions. Potential contaminants were extracted from the analysis. FASTQ formatted files containing DNA sequencing reads were submitted to the short-read correction tool of SOAP*denovo* (Release 1.05), especially designed to correct Illumina GA reads for large plant and animal genomes [20]. The KmerFreq and ErrorCorrection routines were ran with default parameters (seed length = 17, quality cutoff = 5). Illumina sequencing adapters and low quality reads were eliminated using the CLC trimmer function (default limit= 0.05) (CLC Genomics Workbench 4.1 software, CLC Bio, Aarhus, Denmark). Error corrected FASTQ files were then submitted to assembly routines performed on CLC Genomics with *de novo* assembly using short reads (76 bp average length), and mixing of paired end reads (both insert sizes and orientations). The bubble size used was automatically defined by the software as 50 bp. Assembly Length Fraction and Similarity parameters were set to 0.5 and 0.8, respectively. Mismatch, deletion and insertion cost parameters were set to 2, 3 and 3, respectively. The k-mer size on CLC Bio assembler was set to 25 bp and the coverage cutoff to 10X.

Sequence assembly was initially attempted with the sequence fraction of short

insert size contigs (>200 bp) using kmer (de Bruijn graph kmer) overlap information in order to assure unambiguous paths of resulting contigs. The default word length parameter was adjusted to 25 on CLC Bio. Overlaps between sequences were depicted by de Bruijn graph structures [21]. The results were compared with the fraction of contigs >500 bp. The efficiencies of sequence assembly using the >200 pb fraction and the >500 bp fraction were then compared. The >500 bp contig fraction was then submitted to scaffolding procedure using MipScaffolder [22].

B. ruziziensis genome size estimation **-** Genome size estimation was obtained by mapping all usable reads from the short insert size library (>200 bp contigs) on the draft *de novo B.ruziziensis* genome assembly. All aligned reads were used to calculate the distribution of 19-mer frequencies in the sequencing reads using a suffix array provided with Tallymer [23]. The peak depth of 19-mer frequency (M=16) in reads is correlated with the real sequencing depth (N), read length (L), and kmer length (K) [20]. Their relationship can be expressed in a formula: $M = N * (L - K + 1)/L$. Genome size estimation can be obtained by dividing the total sequence length by the real sequencing depth.

In order to assess potential differences between the estimated genome size and the linear size of the draft *de novo* assembly, we performed a *de novo* recognition of repetitive sequences in the draft assembly sequences using the kmer coverage from Tallymer. The genome assembly sequences were queried against the kmer coverage suffixes, and all sequences that had 19-mer occurrences greater than the peak depth were extracted. Segments were grouped based on their coordinates. Using this procedure to evaluate highly repetitive regions in the assembly, we have identified the number of contiguous sequences containing high 19-mer frequencies, and their respective base pair coverage. It follows that the estimated size of the covered genome region using aligned reads is roughly equal to the number of bases assembled plus the size of the draft *de novo* genome assembly. The difference regarding the estimated genome size and the draft *de novo* genome assembly can then be computed.

Gene space metrics and homology-based annotation - *B. ruziziensis de novo* scaffolds were used as references for RNA-Seq sequence data alignment in order to obtain genomic segments to be submitted to gene prediction analytical tools. RNA-seq data from *Brachiaria brizantha*, a species of major economic importance and closely related to *B. ruziziensis*, was kindly provided by Dr. Vera Carneiro. Gene structure models for *Brachiaria* were initially obtained with the spliced aligners programs Tophat

[24] and PASA [25]. Three other eukaryotic gene predictor softwares were also used in this analysis, including SNAP [26], GlimmerHMM [27] and Genemark_ES [28]. All five datasets, containing genome based coordinates depicting possible gene structures were converted to GFF files and used as input to the Evidence Modeler pipeline [29]. Evidence Modeler was configured using different weight values set up to combine *ab initio* predictions, homology-based predictions and transcripts to genome alignments. The purpose was to generate a nearly non-redundant high quality reference gene set for protein-coding gene annotation, gene space coverage metrics and estimation of mRNA abundance for *B. ruziziensis*.

Homology-based prediction was performed querying protein sequences from grasses members of PACMAD clade against the reference gene set by genBlastA. Annotations were loaded into a MySQL database used to distinguish markers located in structural and genomic regions.

Sinteny with sequenced genomes - genBlastA and genBlastG were applied to reveal homologies with the rice genome. A *blastn* analysis was performed querying the predicted *B. ruziziensis* gene sequences against the rice genome (MSU release 7) using genBlastA. At the protein level similarity analysis genBlastG was used to query rice proteins against the *B. ruziziensis* draft genome assembly. Information about the number of bases expanded in the aligned regions, coverage of the alignments and the percentage of identity were extracted and used to generate GFF annotations to explore similarities with the rice genome.

Gene Ontology classification and annotation - The >200 bp contig fraction of the *de novo B. ruziziensis* assembly was blasted against the *Oryza sativa* v6.1 database containing 56,797 rice gene annotations, downloaded from the Michigan State University (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/). Blast hits >200 bp and e-value < 10^{-20} were identified by their Gene Ontology (GO) terms using GOSlim Id. These identifications were submitted to the *Categorizer Ontology Classification* web based software [30] for preliminary analysis (classification method: Plant_GOslim; counting method: single).

The Gene Ontology annotation using the putative gene sequences extracted from the 106,442 scaffolds was submitted to the PFAM database containing signature annotations. The most abundant PFAM domains (those present in 100 genes or more) were detected and classified.

Identification of Simple Sequence Repeat (SSRs) loci and development of microsatellite markers – The draft *de novo* sequence assembly was submitted to simple sequence repeat loci identification in the *B. ruziziensis* genome using PHOBOS (http://www.rub.de/spezzoo/cm/cm_phobos.htm). This analysis was performed in order to obtain a database of microsatellites to be used in the genetic analysis of *B. ruziziensis*. Initially, the location and number of di-, tri-, and tetra-nucleotide SSRs in the draft *de novo* genome assembly was identified and quantified. Then, only contigs with at least 20X coverage, and at least 30% paired end reads were selected for the detection of microsatellite sequences. Finally, a subset of SSRs with five or more di-nucleotide repeats and three or more tri- and tetra-nucleotide repeats was selected.

Estimates of retrotransposon and DNA transposon content in the *B. ruziziensis* genome – The draft de novo sequence assembly was submitted to retrotransposon and DNA transposon sequence analysis using RepeatMasker v. 2.2.23+ [31]. Repbase (v. 20110419) was used as the reference database of eukaryotic repetitive elements [32] using rice (*Oryza sativa L.*) as the query species. Since there is no complete reference genome for *B. ruziziensis*, the selection of rice as a model grass species seemed an appropriate choice. Estimates of retrotransposon and transposon content in the draft of the ruzigrass genome were based on the analysis of three contig fractions: (a) >200 bp contig fraction; (b) >500 bp contig fraction and (c) >2300 scaffold fraction. The incremental variation in contig size was used to verify the potential impact of fractions of different scaffold sizes in Transposable Elements (TE) prediction. Selection of rice as the query species is justified by the amount and quality of reference genomic data available. Also, in order to verify if the retrotransposon and transposon content estimates were reasonable with a *de novo* partial draft of the ruzigrass genome, a recently obtained database of Illumina paired end sequences (Ferreira et al., 2011) of the tropical japonica rice cultivar Chorinho was submitted to the *de novo* genome assembly genome using the same assembly parameters and the same contig size fractions (>200 pb; >500 pb and >2300 pb) used for ruzigrass in the present work. The Nipponbare reference rice genome assembly (v. 6.1 MSU) was used as control. All sequence datasets of ruzigrass and rice were submitted to the same process of TE detection and results were stored in the GFF file format. Low complexity sequences and simple repeats were excluded from the analysis.

Results and Discussion

NGS and *de novo* sequence assembly of the ruzigrass genome - Sequence assembly was based on 265,934,348 DNA short read sequences with 76 bp average length, comprising a database of 20,211,010,488 bp, which corresponds to ~33x coverage of the ruzigrass genome, assuming a genome size of 615 Mbp [5]. Less than 0.10% (200 Kbp) of the sequence assembly genome where detected with some kind of potential DNA contaminants (chloroplast, mitochondrial and exogenous DNA). Assembly was performed using the CLC Assembler (CLC Bio, Aarhus, Denmark), followed by a scaffolding procedure by MipScaffolder [21]. The assembly metrics are presented in Table 1.

Genome assembly was initiated with a dataset of contigs >200 bp (Table 1), but an increase in assembly efficiency was observed when the fraction of contigs >500 pb was considered for analysis. The number of contigs decreased from 280,739 in the >200 bb fraction to 128,020 in the >500 bp fraction. However, the average contig size almost doubled when using the latter fraction (varying from 964 to 1,753) and, most importantly, N50 increased from 1,883 to 2,439 bp. After scaffolding the fraction of contigs >500 bp, the average contig size again increased to 2,047 bp and N50 reached 3,063 bp. Therefore, genome assembly was focused on the >500 bp contig fraction. It is interesting to notice that the percentage of contigs >1 Kbp increased from 74,007 out of 280,739 (26.36%) in the fraction of contigs >200 bp to 66,365 out of 106,442 (62.35%) in the fraction >500 bp after scaffolding (Table 1). The number of scaffolds greater than 10 Kbp almost duplicated, while the number of the contigs <1 Kbp reduced. Therefore, selecting only contigs >500 bp resulted in greater analytical efficiency, culminating in ~218 Mpb assembly of the ruzigrass genome (Table 1).

The final sequence mapping was based on ~32% of the total paired end reads (83,554,104 out of 265,934,348 reads), with a contig coverage of mapped reads of ~28x. The scaffolding process optimized the use of mapped sequences from 4,821,854,656 to 6,090,816,249 (~20.83%) (Table 1). The ~218 Mbp genome draft assembly corresponds to about one third of the estimated *B. ruziziensis* genome size (~615 Mbp - [5].

Assembly efficiency was enhanced by the use of the fraction of contigs >500 bp, which resulted in 6,090,816,249 bp mapped. However, this implied that the majority (69.89%) of the 20,211,010,488 bp input data could not be used in the final genome

assembly (Table 1). A blast of contigs <500 pb against the ~218 Mbp ruzigrass genome draft was performed in order to check if this fraction included redundant reads. Also, mapping the <500 contig fraction on the ruzigrass genome draft using the same stringency and parameters set during the initial mapping procedure could also reveal the nature of the reads in this fraction. The results indicated that 55% of the sequences significantly matched with regions of the ruzigrass genome draft. Also, approximately 50% of the <500 pb contig fraction mapped on draft. Together these results clearly indicate that about half of the contigs belonging to the <500 pb fraction are indeed redundant and already present in the genome draft. The remnant contigs of this fraction could not be used in the final assembly probably due to limitations of the assembly procedure.

Table 1 - *B. ruziziensis* genome assembly metrics. Assembly was initially based on >200 pb and >500 bp contig database fraction, followed by scaffold analysis of >500 pb contig fraction. The total number of paired end reads considered in the analysis was 265,934,348, adding up to 20,211,010,488 bp sequenced.

Genome assembly	>200 bp contig fraction	>500 bp contig fraction	>500 bp contig fraction scaffolding
#mapped reads	72,431,048	63,445,456	83,554,104
# mapped reads (bp)	5,504,759,648	4,821,854,656	6,090,816,249
# contigs	280,739	128,020	106,442
# contigs >500bp	128,020	128,020	106,442
# contigs >1kbp.	74,007	74,007	66,365
# contigs >5 kbp.	6,539	6,539	8,557
# contigs >10kbp.	692	692	1,033
Contig coverage (bp)	270,743,825	224,406,232	217,932,865
Average contig size (bp)	964	1,753	2,047
Contig coverage of mapped reads (%)	20.33	21.49	27.9
Maximum contig size	57,462	57,462	52,016
N50	1,883	2,439	3,063

Test of contig uniqueness in the assembly – In order to verify the uniqueness of the contigs obtained, a mapping analysis of a subgroup of 40,077 <1.0 Kbp contigs on the subgroup of 66,365 >1.0 Kbp contigs was performed using a 0.50 length fraction and 0.25 sequence similarity. The results indicated no sequence matching in the range of 250 to 500 bp with sequence similarity above 25%, indicating the uniqueness of >1.0 Kbp contigs. A Blast analysis of the subgroup of 40,077 <1.0 Kbp contigs on the subgroup of 66,365 contigs >1.0 Kbp indicated a total length of only 4.658.988 bp of the best hit sequences, i. e., only 2.14% of the ~218 Mbp genome draft assembly.

Genome size and coverage of the ruzigrass genome – The results of the *de novo* assembly of the ruzigrass genome indicated an estimated draft size of 286 Mbp. The observed difference between the genome expected size (G = 286 Mbp) and the actual linear size sequence assembly resulting from the unique 106.442 contigs >500 bp (G = 218 Mb, Table 1) could be probably due to the presence of highly repetitive sequences in the ruzigrass genome. It is possible that the assembler tools collapsed the short read fragments in the highly repetitive regions which resulted in a shortening of the linear coverage expected. In order to assess the possible differences regarding the estimated genome size and the actual linear size sequence assembly we performed a *de novo* recognition of repetitive sequences in the draft assembly sequences using the k-mer

coverage from Tallymer [22]. This analysis was done by querying the genome assembly sequences against the k-mer coverage suffixes. After, we extracted all sequences which had 19-mer occurrence greater than the peak depth and performed a sum of the segments based in their coordinates (Figure 1). As a result, we have detected 41,020 contiguous sequences containing highly 19-mer counting, amounting to 57,883,981 of bases in the sequences. It follows that the estimated size of the covered genome region using aligned reads is about the number of bases assembled (217,932,865) plus the number of bases of the 19-mer counting (57,883,981), amounting to a difference regarding the estimated genome size (G = 286 Mbp) of only 3.5%. Therefore, the difference between the genome expected size (G = 286 Mbp) and the actual linear size sequence assembly can be attributed to the presence of highly repetitive sequences in the ruzigrass genome. Previous analysis of the ruzigrass genome by flow cytometry estimated a genome size of 615 Mbp [5]. Based on this size, we estimate a *de novo* genome assembly covering ~35% of the ruzigrass genome.

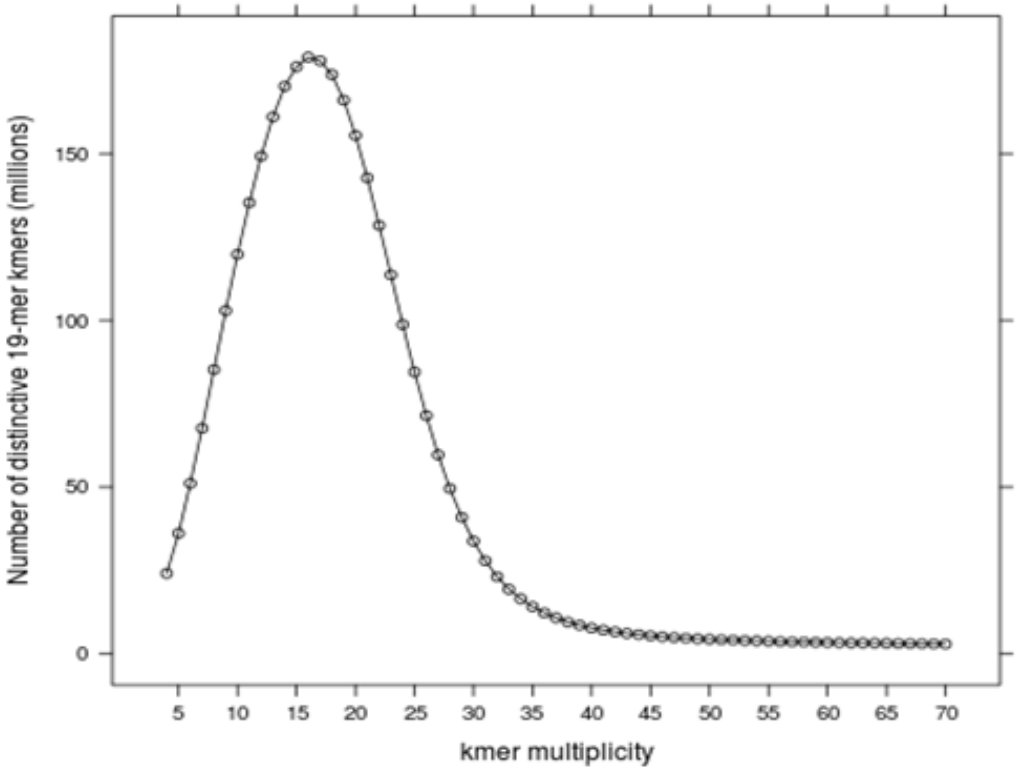


Figure 1 – Distribution of k-mer coverage suffixes of the ruzigrass genome for the extraction of sequences with 19-mer occurrences.

Distribution of contig length in the assembly – The efficiency of the assembly procedure could also be visualized by the cumulative distribution by length of contigs obtained in the analysis. For example, the distribution of the 280,739 contigs belonging

to the >200 bp fraction (Table 1) show that the first 100,000 contigs covered ~200 Mbp of the genome assembly (Figure 2). This represented ~74% of the total contig coverage with this fraction (270,743,824 bp). After the scaffolding procedure of the >500 bp fraction, the first 100,000 larger contigs covered ~218 Mbp, the actual result of the linear size sequence assembly.

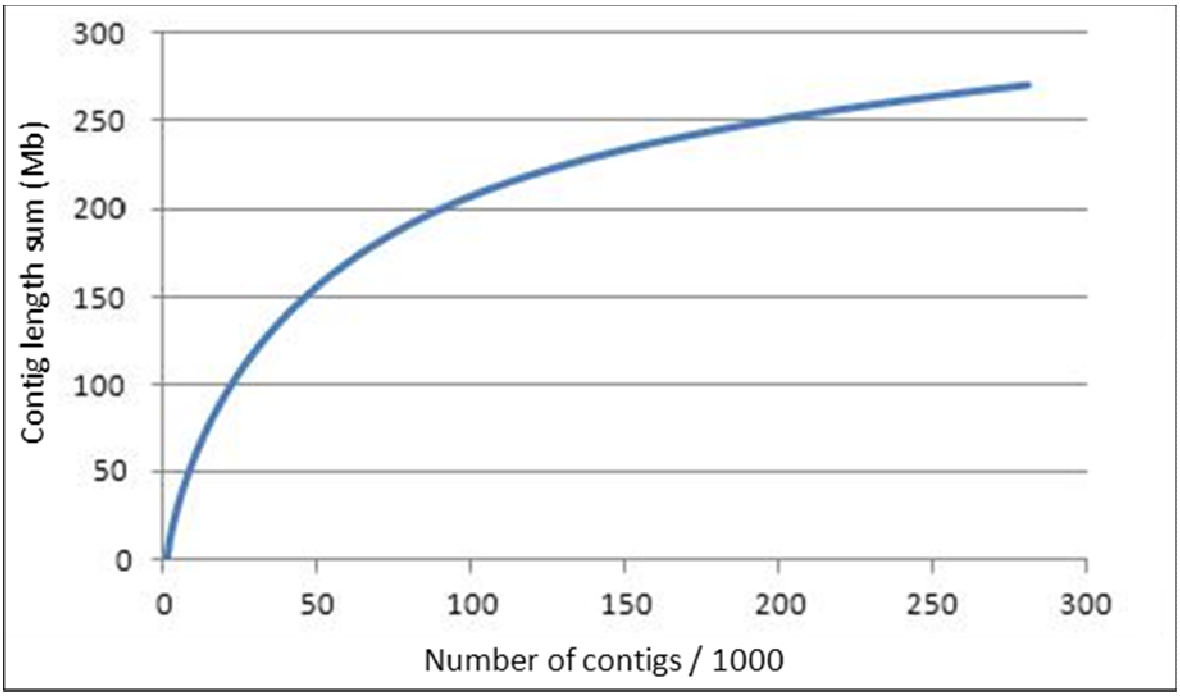


Figure 2 – Cumulative distribution by length of contigs belonging to different contig fractions and their observed genome coverage (y axis = cumulative sum of contig length of contig fraction > 200 bp, in Mpb; x axis = the number of contigs assembled / 1000).

***Brachiaria* spp gene space metrics and homology-based annotation –** *B. ruziziensis* gene prediction and gene content estimates were based on the analysis of the *de novo* sequence assembly of genomic data reported here in combination with transcriptome sequencing data of ovaries at megasporogenesis and megagametogenesis from sexual and apomictic *B. brizantha* accessions [33]. We observed a high rate of significant alignments after mapping the *B. brizantha* sequenced mRNA reads on the *B. ruziziensis de novo* sequence assembly. Sets of putative genes were firstly obtained by combining RNA-Seq data [33] and the *de novo* draft of *Brachiaria ruziziensis* by *ab-initio* predictions, homology-based predictions, transcripts to genome alignments and estimation of mRNA levels. All resulting gene sets were merged using

EvidenceModeler to create a nearly non-redundant high quality reference gene set containing 22,554 sequences with a combined length of 33,919,177 bp (median: 1,047bp; mean: 1,504bp). Then, the *Brachiaria* reference gene set was used to query protein sequences of grass species of PACMAD clade by genBlastA. The reference gene set containing 22,554 target *Brachiaria* sequences was queried against maize, sorghum and switch grass gene sequences (Figure 3). The results allowed for the identification of reliable gene structures - complete or partial – against the target *Brachiaria* sequences. A total of 17,245 gene orthologs were identified between the *B. ruziziensis* gene set and the three grass species. To our knowledge this is the first well-defined subset of possible gene orthologs of the genus *Brachiaria* reported so far.

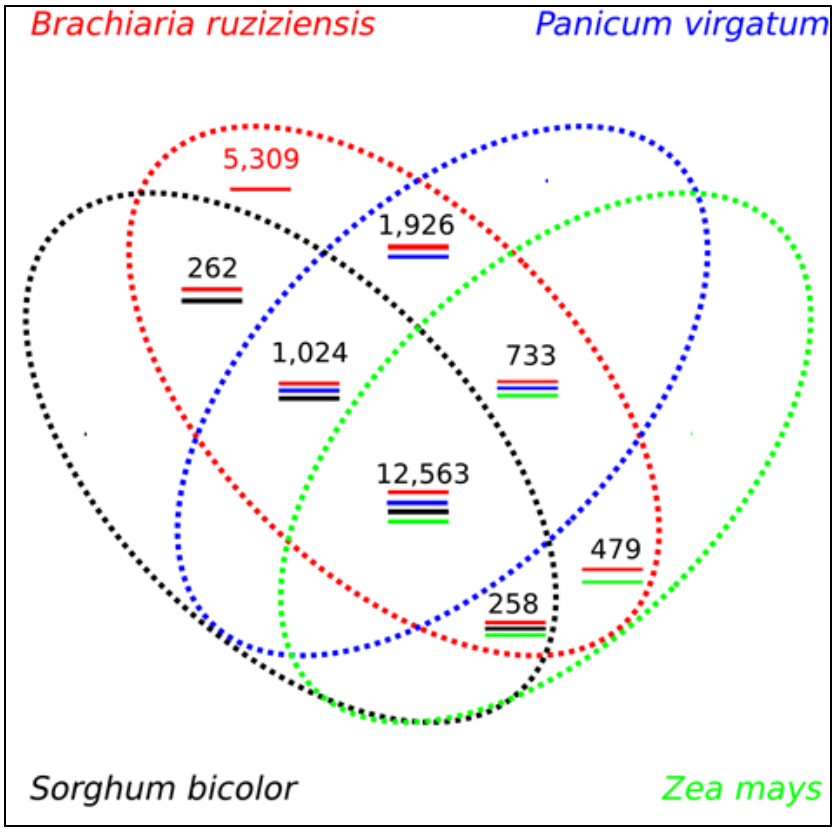


Figure 3 – A database of 22,554 target *Brachiaria* sequences was used to query maize, sorghum and switch grass gene sequence databases. A total of 17,245 common gene orthologs identified between of *Brachiaria* and the three other grass species are depicted.

Statistics of the completeness of the reference gene set based on the CEGMA pipeline shows that 74.60% of the Core Eukaryotic Genes (CEGs) have been mapped on our reference dataset, what provides an estimate of the gene space in the *Brachiaria* genome covered with the combined genomic and mRNA data. Additionally, following CEGMA metrics about completeness and 'paralogy indexes' - referred to as the

proportion of paralog genes in the genome of higher plants (assumed ranging from 51.6% up to 71.3%) - we conservatively estimate that the protein-coding genes in the surveyed genome comprehend between 42,876 (=22,554 x (1+ (1–0.746)) x 1.516) and 49,381 (= 22,554 x(1+(1–0.746)) x 1.713) genes, although it should be noted that this estimate could change depending on ploidy levels, gene splicing behavior and variations on the mating system of *Brachiaria*. For comparison purposes, the latest sequenced maize database (Release 5b) describes a reliable gene set containing 39,656 sequences excluding transposons, pseudogenes, contaminants, and other low-confidence annotations (www.maizesequence.org). Also, the MSU Rice Genome Annotation (Release 7) contains about 39,045 genes excluding pseudogenes and transposable-element related genes (http://rice.plantbiology.msu.edu). It seems that there is a great similarity of genomic space occupied by genes in *Brachiaria ruziziensis* and these other related grass species.

Brachiaria **gene ontology classification** - A total of 280,739 contigs greater than 200 bp length was selected from *de novo* assembly of *B. ruziziensis* to blast against the OSGBD (www.plantgbd.org/OSGBD) data bank containing 56,797 rice gene annotations (Table 2). The 17.023 best blast hits greater than 200 bp, e-value < 10^{-20} and average coverage of 15%, were identified by its Gene Ontology Identification (GOSlim Id) based on the annotation ontology classification on the OSGBD. These annotations brought a list of GO terms linked to this gene model. About 19% of classified transcripts are related to transposable elements. These 88,386 GO terms identifications were submitted to the *Categorizer Ontology Classification* web based software [29] for preliminary analysis and resulted in 314 ontology elements classified in 10 classes summarized in Table 3. The 127 transcripts which compose the "other" group belong to classes under 10 counts from the total list of 314 Gene Ontology groups.

Table 2 - Blast results of *B. ruziziensis* draft genome sequences against *Oryza sativa* cv. Nipponbare transcripts (www.plantgbd.org/OSGBD). Only the best blast hits are reported (>200 bp; e-value < 10e-20; average coverage of 15%).

Number of OSGDB genes	56,797	%
Number of blast hits greater than 200 bp and e-value < -20	17,023	29.97%
Total length of the OSGDB genes bp	75,507,199	
Total length of the greatest e-value hits bp	8,174,867	10.83%
Average gene length coverage (length of hit /length of annotation)	15.00%	
# of genes covering more than 30% (greatest hit length/gene length)	1,938	11.38%
# of genes covering more than 15% (greatest hit length/gene length)	5,604	33.13%

Table 3 - Result counts of Gene Ontology classification distribution of the transcripts identified in the *B. ruziziensis* data set submitted to the Categorizer Ontology Classification system.

Go Class ID	**Definitions**	**Counts**	**%**
GO:000815	Biological process	45	14.33%
GO:0003674	Molecular function	26	8.28%
GO:000575	Cellular component	25	7.96%
GO:0005623	Cell	22	7.01%
GO:0005622	Intracelular	17	5.41%
GO:0009987	Cellular process	16	5.10%
GO:0008152	Matebolic process	14	4.46%
GO:0005488	Binding	12	3.82%
GO:0005737	Cytoplasm	10	3.18%
	Other	127	40.45%

A functional approach based on the PFAM signature of Gene Ontology (GO) annotation using the putative gene sequences resulted in 13,899 terms of which 2,802 were unique. The PFAM signature then identified 13,308 domain annotations of these putative genes, and 2,599 of them were unique. The most abundant PFAM domain observed in this data set, present in more than 100 genes, are listed in Table 4. The number of genes with repeat type signatures represents ~24%.

Table 4 - Most abundant PFAM signature domains found in the *B. ruziziensis* putative gene dataset.

Signature Accession	PFAM Signature (Name)	InterPro Entry Accession	InterPro Entry Name	InterPro Entry Type	Number of genes containing signature	% of genes containing signature
PF00067	p450	IPR001128	Cytochrome P450	Family	270	10.39
PF00069	Pkinase	IPR000719	Proteinkinase, catalyticdomain	Domain	458	17.62
PF00078	RVT_1	IPR000477	Reverse transcriptase	Domain	106	4.08
PF00097	zf-C3HC4	IPR018957	Zinc finger, C3HC4 RING-type	Domain	114	4.39
PF00400	WD40	IPR001680	WD40 repeat	Repeat	112	4.31
PF00560	LRR_1	IPR001611	Leucine-richrepeat	Repeat	194	7.46
PF00646	F-box	IPR001810	F-box domain, cyclin-like	Domain	234	9.00
PF00651	BTB	IPR013069	BTB/POZ	Domain	134	5.16
PF00931	NB-ARC	IPR002182	NB-ARC	Domain	125	4.81
PF01535	PPR	IPR002885	Pentatricopeptiderepeat	Repeat	341	13.12
PF07714	Pkinase_Tyr	IPR001245	Serine-threonine/tyrosine-protein kinase catalytic domain	Domain	271	10.43
PF07727	RVT_2	IPR013103	Reverse transcriptase, RNA-dependent DNA polymerase	Domain	105	4.04
PF08263	LRRNT_2	IPR013210	Leucine-rich repeat-containing N-terminal, type 2	Domain	135	5.19

Simple Sequence Repeats - A total of 430,846 di-, tri- and tetra-nucleotide simple sequence repeats (SSRs) was identified and annotated on the *B. ruziziensis* genome scaffolds (Table 5). Tri-nucleotide repeats were the most abundant class of microsatellites observed (49%), which is consistent with the findings in other close grass genomes, such as *Panicum virgatum* [8]. A subset of 200,873 di-, tri-, tetranucleotide SSRs with integer number of motif repeats ('perfect' microsatellites) was selected for further analysis. Selection criteria for this analysis included a minimum number of five repeats for di-nucleotide motifs and three repeats for tri- and tetra-nucleotides, resulting in 147.870 perfect SSRs. This selection reduced strongly the percentage of di-nucleotide relative to tri-nucleotide and tetra-nucleotide microsatellites (Table 5). After this selection, the most frequent simple sequence repeat motifs observed was CCG (29,235) comprising 19.77% of the microsatellites, followed by AGC (14,552) and AGG (14,064). Among di-nucleotide SSRs, AG (4,695), AT (3,412) and AC (2,728) were the most abundant motifs; and among the tetra-nucleotides, AAAT (1,062), ATGC (998) and AAAG (897) were the most common ones (Table 5). A subset of microsatellites was further selected to develop new markers for genetic analysis of *B. ruziziensis*. Only microsatellite loci with a minimum 20x coverage and at

least 30% of aligned paired end sequences at the microsatellite locus were selected. A total of 18,162 SSRs mapped on 8,671 contigs was selected. It was observed that the proportion of di- and tetra-nucleotide SSR motifs did not vary during the pipeline of selection criteria (Table 5), except for 2,541 perfect microsatellite sequences found within predicted gene regions. The 18,162 new SSRs selected in the present work are being compared with a recent set of microsatellite markers obtained from single end Illumina reads and used for genetic analysis of *B. ruziziensis* [6].

Table 5 - SSRs annotation of di-, tri- and tetra-nucleotide repeats of the *B. ruziziensis* genome.

	Total SSRs	**Perfect SSRs**	**Selection Criteria***	**Coverage > 20x****	**Paired end reads>30%****	**# SSR in predicted genes**
DI	96,458 (0.22)	65,701 (0.33)	12,698 (0.09)	2,495 (0.09)	1,603 (0.09)	89 (0.03)
TRI	211,671 (0.49)	122,848 (0.61)	122,848 (0.83)	24,087 (0.83)	15,061 (0.83)	2,327 (0.92)
TETRA	122,717 (0.28)	12,324 (0.06)	12,324 (0.08)	2,309 (0.08)	1,530 (0.08)	125 (0.05)
Total	430,846	200,873	147,870	28,891	18,162	2,541

*Minimum number of motif repeats: di≥5; tri≥ 3; tetra≥3;
**Percentage of paired end reads at the microsatellite locus ≥ 0.30
Numbers in parenthesis represent the relative percentage of SSR motif class.

Transposable Elements (TE) in the *B. ruziziensis* genome - Repetitive DNA, including retrotransposons and DNA transposons, comprised only 3.53% of the total sequences assembled in ~218 Mbp of the ruzigrass genome (Table 6). This fraction of repetitive sequences does not include low-complexity sequences, such as microsatellites. Classification of the observed transposable elements into known classes revealed that the majority of repetitive sequences is composed of retrotransposons (2.64%). Only 0.89% of the transposable elements were DNA transposons (Table 6). The most abundant repeats identified are long-terminal repeat elements (1.60%), followed by Gypsy-type elements (0.80%) and 0.79% Copia-type elements (0.79%).

In order to check the veracity of the apparent low percentage of transposable elements in the de novo assembly of the *B. ruziziensis* genome, we compared estimates of repetitive sequence content in three different ruzigrass genome databases (de novo assembly, based on scaffolding of >200 bp contigs; de novo assembly, based on scaffolding of >500 bp contigs, and *de novo* assembly, based on scaffolding of >2300 bp contigs) with *de novo* assemblies of the rice genome using on the same assembly parameters adopted for ruzigrass (Table 6). The logic here was to check if a *de novo*

genome assembly of a species of similar size such as rice using the same methodology used for ruzigrass could reveal a bias towards low percentage of transposable elements in the genome assembly. The Nipponbare reference rice genome assembly (MSU release 7) was used as control.

The results, as expected, show that the de novo assemblies of the rice genome indeed cause a significant reduction on the TE content estimate (Table 6). This is probably due to the inherent difficulties of de novo assembling based on short sequencing reads to deal with repetitive sequences [11]. The largest variation of TE predictions between de novo and reference assemblies happens with longer TE, such as LTR elements (Table 6).

The data, however, reveals that the TE content in the *B. ruziziensis* genome seems to be lower than in rice and other grass species. Retroelements, for instance, cover approximately 22.43% of the rice genome, as observed in the rice reference genome MSU v. 6.1 (Table 6). The estimates of retroelement coverage on the three *de novo* assemblies of the rice genome varied from 3.99 to 6.05%. This indicates that the retroelement content estimates on the rice *de novo* assemblies varies from only 17.79% to 26.97% of the retroelement content described on the rice reference genome. In other words, three *de novo* assemblies of the rice genome could identify only 17.79% to 26.97% of the retroelements found in the rice genome. Since the assembly parameters and methodology used for these three *de novo* assemblies were the same employed on the three genome assemblies of the *B. ruziziensis* genome, this would imply that the retroelement content of the *B. ruziziensis* would vary from 11.11% to 12.83%, which is approximately half of the retroelement content observed in the rice genome.

Similarly, the estimates of DNA transposons coverage on the three *de novo* assemblies of the rice genome varied from 7.15 to 9.20% (Table 6). This indicates the DNA transposon content estimates on the rice *de novo* assemblies varies from 50.57% to 65.06% of the DNA transposon content described on the rice reference genome. Thus, DNA transposon content of *B. ruziziensis* would vary from 1.38% to 1.94%, which is approximately 10x smaller than the DNA transposon content (14.14%) observed in the rice genome (Table 6).

Considering the different classes of TE, the TE content estimate of the *B. ruziziensis* genome would vary from 8.96 to 11.32%, which is smaller than the total TE

content observed in rice reference genome (36.57%). The *B. ruziziensis* genome, therefore, seems to have a smaller TE content than the rice genome.

In order to observe the relationship between TE and gene space distribution in *B. ruziziensis* genome, the contigs with and without gene annotations were identified and a TE analysis was carried out on this data set. A total of 62.168 repetitive elements on annotated contigs covering 9,290,875 bp, which represents only 4.31% of the de novo genome assembly. TE annotated on gene space covered only 418,057 bp, roughly 4% of the gene space. Therefore, the results suggest a small TE presence on the *B.ruziziensis* gene space.

Table 6 – Percent estimate of Transposable Elements (TE) coverage of three de novo assemblies of ruzigrass (*B. ruziziensis*) and rice (*Oryza sativa*) genomes, after classification of elements on different TE classes.

Contig minimum size	**Contig minimum size**	**# Scaffolds**	**Retroelements**	**SINEs**	**LINEs**	**LTR**	**Ty1/Copia**	**Gypsy/DI**	**DNA transposons**	**Hobo Activator**	**Tc1-IS630-Pogo**	**n-Spm**	**MuDR-IS905**	**Tourist/Harbinger**	**Total**
***B. ruziziensis* (*de novo* assembly)**	**200 bp**	**280,739**	**3.46**	0.17	1.03	2.26	0.98	1.27	**1.26**	0.11	0.3	0.18	0.35	0.32	**4.72**
***B. ruziziensis* (*de novo* assembly)**	**500 bp**	**106,442**	**2.64**	0.11	0.93	1.60	0.79	0.80	**0.89**	0.09	0.18	0.12	0.29	0.18	**3.53**
***B. ruziziensis* (*de novo* assembly)**	**2300 bp**	**29,511**	**2.14**	0.08	0.85	1.21	0.51	0.69	**0.70**	0.06	0.17	0.08	0.24	0.13	**2.84**
***O. sativa* cv. Chorinho (*de novo* assembly)**	**200 bp**	**186,502**	**6.05**	0.41	0.93	4.71	0.94	3.62	**9.20**	0.38	2.13	0.90	2.20	2.06	**15.25**
***O. sativa* cv. Chorinho (*de novo* assembly)**	**500 bp**	**102,304**	**5.33**	0.41	0.95	3.97	0.86	2.98	**9.07**	0.37	2.14	0.84	2.16	2.04	**14.4**
***O. sativa* cv. Chorinho (*de novo* assembly)**	**2300 bp**	**30,845**	**3.99**	0.33	0.97	2.70	0.71	1.91	**7.15**	0.30	1.76	0.62	1.73	1.55	**11.14**
***Oryza sativa* cv. Nipponbare v.6.1 MSU**	**12 chr***	**12 chr***	**22.43**	0.39	0.89	21.16	3.03	17.67	**14.14**	0.55	2.42	3.46	3.24	2.49	**36.57**

**chr = chromosome*

Conclusion

1. Sequence assembly of the *B. ruziziensis* genome was based on 265,934,348 Illumina DNA short read sequences, comprising a database of 20,211,010,488 bp, which corresponds to ~33x coverage of the ruzigrass genome, assuming a genome size of 615 Mbp.
2. During assembly, selecting only the fraction of contigs >500 pb resulted in greater analytical efficiency, culminating in ~218 Mpb draft of the ruzigrass genome.
3. The final sequence mapping was based on ~32% of the total paired end reads obtained (83,554,104 out of 265,934,348 reads), with a contig coverage of mapped reads of ~28x.
4. The ~218 Mbp genome draft assembly corresponds to about 35% of the estimated *B. ruziziensis* genome size.
5. Assembly efficiency was enhanced by the use of the fraction of contigs >500 bp, which resulted in 6,090,816,249 bp mapped. However, this implied that the majority (69.89%) of the 20,211,010,488 bp input data could not be used in the final genome assembly. It was observed that about half of the contigs belonging to the <500 pb fraction are indeed redundant and already present in the genome draft.
6. The results of the *de novo* assembly of the ruzigrass genome indicated an estimated draft size of 286 Mbp. Therefore, the difference between the genome expected size (G = 286 Mbp) and the actual linear size sequence assembly can be attributed to the presence of highly repetitive sequences in the ruzigrass genome.
7. Considering the different classes of Transposable Elements (TE), the TE content estimate of the *B. ruziziensis* genome would vary from 8.96 to 11.32%, which is smaller than the total TE content observed in rice reference genome (36.57%). The *B. ruziziensis* genome, therefore, seems to have a smaller TE content than the rice genome. The results suggest a small TE presence on the *B.ruziziensis* gene space.
8. A nearly non-redundant high quality reference gene set of the *B. ruziziensis* genome was obtained, which contain 22,554 sequences with a combined length of 33,919,177 bp (median: 1,047bp; mean: 1,504bp). A total of 17,245 gene orthologs were identified between the *B. ruziziensis* gene set and the three grass

species (sorghum, maize and switch grass). It seems that there is a great similarity of genomic space occupied by genes in *Brachiaria ruziziensis* and theses species. To our knowledge this is the first well-defined subset of possible gene orthologs of the genus *Brachiaria* reported so far.

9. The estimate of the protein-coding genes in the surveyed genome comprehend between 42,876 and 49,381 genes, although it should be noted that this estimate could change depending on ploidy levels, gene splicing behavior and variations on the mating system of *Brachiaria*.
10. A total of 430,846 di-, tri- and tetra-nucleotide simple sequence repeats (SSRs) was identified and annotated on the *B. ruziziensis* genome scaffolds. Tri-nucleotide repeats were the most abundant class of microsatellites observed. A set of 18,162 perfect new SSRs was selected for use in genetic analysis and breeding of *B. ruziziensis.*
11. Millions of DNA sequence reads obtained in a single run of NGS equipment provided enough data to initiate the genomic analysis of *B. ruziziensis*.

ACKNOWLEDGEMENTS

We would like to thank Fausto Souza Sobrinho for providing the *Brachiaria ruziziensis* accession *FSS-1* used in this work. Our thanks to Dr. Vera Carneiro and colleagues for kindly providing access to transcriptome sequencing data of ovaries from sexual and apomictic *B. brizantha* accessions. This research was sponsored by EMBRAPA Macroprograma 2 – Grant # 02.12.02.002.00.00.

References

1. Keller-Grein, G., B.L. Maass, and J. Hanson, *Natural variation in Brachiaria and existing germplasm collections.* CIAT publication; no. 259, 1996.
2. BARCELLOS, A.d.O.A., R.P. de; KARIA, C.T.; VILELA, L., *Potencial e uso de leguminosas forrageiras dos gêneros Stylosanthes, Arachis e Leucaena.* . SIMPÓSIO SOBRE MANEJO DA PASTAGEM, 17 2001. **17**.
3. Lapointe, S. and J. Miles, *Germplasm case study: Brachiaria species.* Pastures for the Tropical Lowlands, CIAT, Cali, Colombia, 1992: p. 43-55.
4. Azevedo, A.L., et al., *High degree of genetic diversity among genotypes of the forage grass Brachiaria ruziziensis (Poaceae) detected with ISSR markers.* Genet Mol Res, 2011. **10**(4): p. 3530-8.
5. Ishigaki, G., et al., *Estimation of genome size in Brachiaria species.* Grassland Science, 2010. **56**(4): p. 240-242.

6. Silva, P.I., et al., *Development and validation of microsatellite markers for Brachiaria ruziziensis obtained by partial genome assembly of Illumina single-end reads.* BMC Genomics, 2013. **14**(1): p. 17.
7. Varshney, R.K., et al., *A comprehensive resource of drought- and salinity-responsive ESTs for gene discovery and marker development in chickpea (Cicer arietinum L.).* BMC Genomics, 2009. **10**: p. 523.
8. Varshney, R.K., et al., *Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers.* Nat Biotechnol, 2012. **30**(1): p. 83-9.
9. Wang, K., et al., *The draft genome of a diploid cotton Gossypium raimondii.* Nat Genet, 2012. **44**(10): p. 1098-1103.
10. Varshney, R.K., et al., *Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement.* Nat Biotechnol, 2013. **31**(3): p. 240-6.
11. Rahman, A.Y.A., et al., *Draft genome sequence of the rubber tree Hevea brasiliensis.* BMC Genomics, 2013. **14**(1): p. 75.
12. Pop, M. and S.L. Salzberg, *Bioinformatics challenges of new sequencing technology.* Trends Genet, 2008. **24**(3): p. 142-9.
13. Pop, M., *Genome assembly reborn: recent computational challenges.* Brief Bioinform, 2009. **10**(4): p. 354-66.
14. Shulaev, V., et al., *The genome of woodland strawberry (Fragaria vesca).* Nat Genet, 2011. **43**(2): p. 109-16.
15. Brenchley, R., et al., *Analysis of the bread wheat genome using whole-genome shotgun sequencing.* Nature, 2012. **491**(7426): p. 705-710.
16. Varshney, R.K., et al., *Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement.* Nat Biotech, 2013. **31**(3): p. 240-246.
17. Argout, X., et al., *The genome of Theobroma cacao.* Nat Genet, 2011. **43**(2): p. 101-108.
18. Doyle, J. and J. Doyle, *A rapid DNA isolation procedure for small quantities of fresh leaf tissue*. 1987.
19. Ferreira, M.E. and D. Grattapaglia, *Introducao ao uso de marcadores moleculares em analise genetica*. Documento / EMBRAPA-CENARGEN;20. 1996, [S.l.]: Ministerio da Agricultura e do Abastecimento [etc.].
20. Li, R., et al., *The sequence and de novo assembly of the giant panda genome.* Nature, 2010. **463**(7279): p. 311-317.
21. Zerbino, D.R. and E. Birney, *Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.* Genome Research, 2008. **18**(5): p. 821-829.
22. Salmela, L., et al., *Fast scaffolding with small independent mixed integer programs.* Bioinformatics, 2011. **27**(23): p. 3259-3265.
23. Kurtz, S., et al., *A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.* BMC Genomics, 2008. **9**(1): p. 517.
24. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-1111.
25. Haas, B.J., et al., *Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.* Nucleic Acids Res, 2003. **31**(19): p. 5654-66.
26. Korf, I., *Gene finding in novel genomes.* BMC Bioinformatics, 2004. **5**(1): p. 59.
27. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Research, 1999. **27**(23): p. 4636-4641.
28. Ter-Hovhannisyan, V., et al., *Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training.* Genome Res, 2008. **18**(12): p.

1979-90.
29. Haas, B.J., *Analysis of alternative splicing in plants with bioinformatics tools.* Curr Top Microbiol Immunol, 2008. **326**: p. 17-37.
30. Zhi-Liang, H., J. Bao, and J. Reecy, *CateGOrizer: a web-based program to batch analyze gene ontology classification categories.* Online J Bioinformatics, 2008. **9**: p. 108-112.
31. Smit, A.F., *The origin of interspersed repeats in the human genome.* Curr Opin Genet Dev, 1996. **6**(6): p. 743-8.
32. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.
33. Silveira, E.D., et al., *Expressed sequence-tag analysis of ovaries of Brachiaria brizantha reveals genes associated with the early steps of embryo sac differentiation of apomictic plants.* Plant Cell Rep, 2012. **31**(2): p. 403-16.

X. CAPÍTULO 3

Sequenciamento, montagem *de novo*, caracterização do genoma de cloroplasto de quatro espécies de *Brachiaria* e desenvolvimento de marcadores para diferenciação de espécies do gênero.

CAPITULO 3

Sequenciamento, montagem e caracterização do genoma cloroplástico (cpDNA) de quatro espécies de *Brachiaria* e desenvolvimento de marcadores indel para diferenciação de espécies do gênero

Resumo

Nas últimas décadas, a análise de polimorfismo de DNA possibilitou uma grande ampliação do conhecimento da filogenia de plantas, particularmente das angiospermas. O sucesso desta análise tem por base, principalmente, a avaliação da variação da estrutura e da sequência nucleotídica do genoma cloroplástico. Alterações microestruturais, tais como pequenas inserções e deleções (indels) do DNA cloroplástico, são úteis para resolver relações filogenéticas entre acessos de um mesmo gênero, para inferir as relações de vínculo genético entre acessos mais relacionados, ou para serem usadas na rápida discriminação de espécies em programas de conservação e uso de germoplasma. Isto pode ser particularmente importante entre acessos de grupos morfologicamente muito semelhantes, onde há grande dificuldade de separação de espécies pela ausência de descritores morfológicos, como ocorre nas espécies de *Brachiaria*. A destacada conservação de tamanho, organização e sequência do genoma cloroplástico justifica o emprego da análise do cpDNA na compreensão da filogenia de espécies de *Brachiaria* e estimula a sua potencial aplicação no desenvolvimento de ferramentas de apoio a programas de conservação e uso de recursos genéticos de espécies deste gênero. O emprego de cpDNA em análise filogenética é favorecido ainda pela facilidade de extração de DNA e abundância de DNA extraído, devido ao grande número de cópias do cpDNA em cada unidade celular.

O sequenciamento do genoma cloroplástico geralmente é feito através da extração e separação do cpDNA do genoma nuclear e mitocondrial, seguido por amplificação e purificação para a construção da bibliotecas. Neste trabalho, optou-se pelo sequenciamento NGS de amostras de DNA total, provendo um alto rendimento de segmentos de leitura do cpDNA. Os resultados permitiram a recuperação de quantidade suficiente de segmentos de leitura exclusivos do cpDNA para a montagem do cpDNA de *B. ruziziensis, B. brizantha, B. decumbens e B. humidicola*. Os quatro genomas de cloroplasto montados possuem uma estrutura circular típica, com uma grande região de cópia única (*Large Single Copy* - LSC) e uma pequena região de cópia única (*Small*

Single Copy - SSC), separadas por duas cópias de inversão repetida (*Inverted Repeat* - IR). O tamanho dos cpDNA obtidos variaram entre 138.765 bp em *B. ruziziensis* e 138.976 bp em *B. humidicola*. O genoma do cloroplasto das quatro espécies de *Brachiaria* contém 118 genes únicos, dos quais 18 são duplicados nas regiões invertidas IRs, perfazendo um total de 136 genes de função conhecida. Além disso, existem nove ORFs e três pseudogenes. A cobertura linear alcançada neste trabalho pelo somatório das sequências montadas *de novo* (*scaffolds*) variou entre 92,89 a 99,45%. O alinhamento das sequências montadas de cpDNA das quatro espécies possibilitou a seleção de regiões indel que permitem a separação de acessos de cada espécie. Foram selecionados para validação um total de 18 indels que apresentam polimorfismo de inserção/deleção *in silico* e permitem distinguir as quatro espécies de *Brachiaria* (*B. ruziziensis, B. brizantha, B. decumbens e B. humidicola*). Análise laboratorial confirmou a utilidade dos marcadores indels na separação de acessos de diferentes espécies de *Brachiaria*. As relações filogenéticas das quatro espécies de *Brachiaria* foram ainda exploradas por uma análise comparativa das quatro sequências completas do cpDNA (usando referência), juntamente com sequências completas de outras gramíneas depositadas no Genbank. Todas as árvores filogenéticas construídas tiveram a mesma topologia e indicam que *B. decumbens* e *B. brizantha* podem ser derivadas de um ancestral comum com *B. ruziziensis.* Indicam ainda que *B. humidicola* é mais distante de *B. ruziziensis, B. decumbens e B. brizantha*. Estima-se que o tempo de divergência entre *B. decumbens* e *B. brizantha* seja de apenas 2,5 MYA, e que estas duas espécies tenham se separado do ancestral que originou *B. ruziziensis* há 14 MYA. Isto provoca a hipótese de que *B. decumbens* e *B. brizantha* tenham surgido do ancestral de *B. ruziziensis* por evento(s) de poliploidização, que isolou reprodutivamente estas espécies.

Introdução

Classificações incorretas de espécies de *Brachiaria* comumente utilizadas nas pastagens tropicais são frequentes no gênero. A falta de descritores morfológicos que permitam a fácil classificação dos acessos, especialmente na fase vegetativa, torna complexa a separação e a intensificação do uso dos acessos. O intercâmbio de germoplasma sem que haja a devida cautela na classificação intensifica certa confusão sobre a identidade dos acessos. Diversos especialistas [1-3] destacam a necessidade de classificar acessos e discriminar espécies corretamente, inclusive para que os bancos de germoplasma possam ser utilizados com eficiência no melhoramento genético de

espécies do gênero.

Nas últimas décadas, a análise de polimorfismo de DNA possibilitou uma grande ampliação do conhecimento da filogenia de plantas, particularmente das angiospermas. O sucesso desta análise tem por base, principalmente, a avaliação da variação da estrutura e da sequência nucleotídica do genoma cloroplástico [4, 5] e da região repetitiva do DNA ribossomal no genoma nuclear [5-7]. A região rDNA do genoma nuclear, composta de repetições em tandem de segmentos (*repeats*) que contêm genes do rRNA e regiões espaçadoras intergênicas, exibe uma taxa de evolução relativamente rápida, que permite uma análise de forma análoga ao cpDNA [8]. O emprego da análise molecular do genoma cloroplástico e da região rDNA do genoma nuclear na filogenia de espécies de braquiária pode, potencialmente, contribuir para discriminação das diferentes espécies deste gênero.

Na maioria das plantas, o cpDNA consiste em um único cromossomo circular, com uma estrutura quadripartida, que inclui uma região grande de cópia única (*Large Single Copy* - LSC) e uma pequena região de cópia única (*Small Single Copy* - SSC), separadas por duas cópias de inversão repetida (*Inverted Repeat* - IR), cada qual com ~25Kbp de comprimento [4, 9]. Nas espécies com menores genomas cloroplásticos, em geral, constata-se a perda de uma das cópias da IR (*Inverted Repeat*), como nas coníferas e em algumas leguminosas. Nas espécies com maior cpDNA (ex. *Pelargonium*) observa-se uma expansão de tamanho da IR, sem que haja alterações de organização e de complexidade [4].

O sequenciamento e alinhamento do genoma cloroplástico de quatro espécies vegetais (tabaco – *Nicotiana tabacum* [10]; *Merchantia polymorpha* [11]; arroz – *Oryza sativa* [12]; *Epifagus virginiana* [13]) possibilitou um grande avanço inicial no conhecimento da estrutura, organização e conteúdo gênico do cpDNA. Alterações de conteúdo gênico, por exemplo, são raras no cpDNA e geralmente associadas a deleções ou pequenas inversões de porções do genoma.

A ordem linear dos genes e arranjos cromossômicos no cpDNA pouco variam nas angiospermas. Chama a atenção, por exemplo, a esmagadora presença da grande inversão repetida (*Inverted Repeat - IR*) no cpDNA das mais diversas espécies vegetais, mesmo aquelas taxonomicamente distantes. Estas características possibilitam utilizar as variações de sequência medidas no cpDNA como um relógio molecular. Isto permite que inferências de distância evolutiva possam ser feitas com base em variações na

sequência do cpDNA em vários níveis taxonômicos. O emprego de cpDNA em análise filogenética é favorecido ainda pela facilidade de extração de DNA e abundância de DNA extraído, devido ao grande número de cópias do cpDNA em cada unidade celular. A herança do genoma nuclear é bi-parental e segue o padrão Mendeliano. Já a herança do genoma cloroplástico é quase sempre materna e segue um padrão clonal, isto é, o cpDNA é herdado como um haplótipo devido a ausência de recombinação. A combinação de variações estruturais e de sequência dos genomas nuclear (ex. rDNA) e cloroplástico, portanto, permitem analisar diferentes componentes da história evolutiva de uma espécie.

Substituições nucleotídicas ocorrem em taxas relativamente baixas nas regiões gênicas do cpDNA assim como em outras regiões deste genoma. A forte seleção sobre maquinaria fotossintética impõe restrições sobre as taxas de mutação de nucleotídeos. Embora haja pressão seletiva para conservar as sequências do genoma do cloroplasto, que são fundamentais para o desenvolvimento do aparato fotossintético, variações de sequência e estrutura podem ser detectadas e usadas em análise filogenética [15]. Uma lista de genes do cpDNA vem sendo comumente usada em análise filogenética, incluindo os genes *psb*A, *psb*D, *psa*B, *psb*B, *psb*C, *psa*A, *rbc*L, *atp*B, *ndh*A, *atp*A, *ndh*D, *rpo*B, *rpo*Cl, *ndh*A, *ndh*F, *rpo*C2, *mat*K. Estes genes foram incialmente selecionados por terem diferentes taxas de substituição de nucleotídeos, serem suficientemente longos (>1kb) e presentes na maior parte das angiospermas. O fato de apresentarem uma taxa de substituição de nucleotídeos bastante variável possibilita análise filogenética em vários níveis taxonômicos.

A seleção de algumas destas regiões do cpDNA para classificação taxonômica de espécies vegetais deu origem ao conceito de "DNA *barcoding*" ou código de barras de DNA em plantas. DNA *barcoding* baseia-se no emprego de uma ou poucas regiões do DNA para distinguir a maioria das espécies do planeta. O ponto de partida para DNA *barcoding* é a construção de um banco de dados de sequências de DNA de várias espécies (e dentro de espécies), analisadas nos genes selecionados, para fazer inferências taxonômicas. Este banco pode servir de apoio também na identificação de "novas espécies", através de comparações entre as sequências depositadas e as sequências de novos acessos coletados. Em animais, onde o conceito foi criado, uma parte do gene da oxidase do citrocromo vem sendo usada com grande eficiência na discriminação de espécies nos últimos 10 anos. Em plantas, o esforço é mais recente e um conjunto de regiões *barcoding* tão eficiente quanto em animais continua a ser

perseguido.

Uma grande utilidade da análise do DNA cloroplástico, demonstrada em inúmeros grupos de espécies com variação de ploidia e de tamanho no DNA nuclear, é capacidade de resolução de relações de vínculo genético e do tempo de surgimento de híbridos interespecíficos e de espécies poliploides. Apesar de ser mais conservado, o cpDNA apresenta diferenças suficientemente grandes e complexas capazes de possibilitar a diferenciação e análise de divergência evolutiva entre espécies. A destacada conservação de tamanho, organização e sequência do genoma cloroplástico justifica o emprego da análise do cpDNA na compreensão da filogenia de espécies de *Brachiaria* e estimula a sua potencial aplicação no desenvolvimento de ferramentas de apoio a programas de conservação e uso de recursos genéticos de espécies deste gênero. Avanços significativos foram alcançados nos últimos anos [16, 17] através do emprego combinado de sequências codificadores do genoma cloroplasto (ex. rbcl, matK e PsbA_TrnH) e da região ITS (*internal transcribed spacers*) do rDNA nuclear (nrDNA ITS).

Este estudo teve os seguintes objetivos: (1) sequenciar e comparar as regiões *barcoding* de plantas (rbcl, matK e PsbA_TrnH) e região ITS (*internal transcribed spacers*) do rDNA nuclear; (2) sequenciar, montar e comparar as sequências completas de cpDNA de quatro espécies de *Brachiaria* (*B. brizantha, B. decumbens, B. ruziziensis e B. humidicola*); (3) comparar os resultados da análise de regiões *barcoding* com os resultados de sequenciamento e montagem completa de cpDNA; (4) estudar as relações filogenéticas entre as quatro espécies com base nas sequências completas de cpDNA obtidas; (5) selecionar, desenvolver e validar marcadores indel (inserção/deleção) do cpDNA para a rápida identificação de espécies de *Brachiaria*.

Material e Métodos

Material Vegetal

Os seguintes acessos de braquiária foram utilizados para a extração de DNA, sequenciamento de cpDNA e análise filogenética neste trabalho: (1) FSS-1, clone de *B.ruziziensis* obtido de autofecundação de planta de população aberta da variedade Kennedy, mantida pelo Programa de Melhoramento Genético, Embrapa Gado de Corte, Juiz de Fora, MG; (2) *B. ruziziensis* acesso 06; (3) *B. ruziziensis* acesso 10; (4) *B. brizantha* cv. Marandú, CIAT accession # 6294 (código neste trabalho: acesso 12); (5) *B. brizantha* cv. Piatã (código neste trabalho: acesso 14); (6) *B. decumbens* cv.

Basilisky, CIAT accession # 606 (código neste trabalho: acesso 18); (7) *B. humidicola* cv. Tupi (código neste trabalho: acesso 19).

Análise de regiões *barcoding* do DNA

Para a análise de variação de sequência de DNA visando a diferenciação de espécies de braquiária foram selecionadas as seguintes regiões *barcoding* do DNA vegetal: (a) uma região do genoma nuclear - região ITS (*internal transcribed spacers 1 e 2, com região central 5.8S*) do rDNA nuclear, [18]; (b) quatro regiões do genoma cloroplástico: a região altamente variável do espaço intergênico *trn*H-*psb*A [19], a região *trn*L (UAA), incluindo o intron e o espaço intergênico entre *trn*L (UAA) 3' exon e o *trn*F (GAA) [20], e partes das regiões dos genes *rbc*L e *mat*K, comumente usadas para *barcoding* em plantas [21].

Para esta análise, o DNA genômico foi purificado [22] e as regiões ITS 1 e ITS 2, juntamente com a região central 5.8S rDNA do rDNA nuclear foi amplificada por PCR, seguindo os procedimento descritos por White et al. (1990) [23]. A região do espaço intergênico trnH-psbA do cpDNA foi amplificada utilizado os primers *psb*A3'f [24] e *trn*Hf [25]. A região *trn*L (UAA) intron e o espaço intergênico entre *trn*L (UAA) 3' exon e o *trn*F (GAA) foram amplificadas em reações separadas com os primers c-d, e e-f, respectivamente [20]. Parte do *rbc*L gene foi amplificado usando o par de primers *rbc*La_f e *rbc*La_r [26], usando o mesmo PCR mix empregado na amplificação *trn*H-*psb*A. A amplificação da região *mat*K foram usados os primers 1R_KIM e 3F_KIM (Ki-Joong Kim, não publicado, listados em: Dunning & Savolainen, 2010)[27], e as condições de PCR foram as mesmas usadas para amplificação do gene *rbc*L.

Os produtos de PCR foram analisados em gel de agarose após eletroforese e preparados para sequenciamento usando o kit PCR ExoSAP (GE Biosciences). As duas fitas dos produtos de PCR foram sequenciadas usando o kit Big Dye v.3.1 (Applied Biosystems) em um sequenciador automatic ABI3700 (Applied Biosystems). As sequências (*foward/reverse*) foram montadas usando o software ChromasPro v1.5 (Technelysium Pty Ltd), e os contigs de cada loco *barcoding* alinhados para análise usando o programa MUSCLE v3.5 [28] e manualmente editados com BioEdit v7.0.9 [29].

Uma análise cladística dos dados de cada loco *barcoding* foi realizada utilizando o critério ML (Maximum Likelihood) com o programa MEGA 5 [30]. O suporte para ramificações (*branch support*) dos dendrogramas foi obtido por 100 pseudo-replicações

bootstrap [31].

Sequenciamento NGS do genoma cloroplástico

O DNA para sequenciamento do cpDNA dos acessos *de B. ruziziensis, B. brizantha, B. decumbens e B. humidicola* foi extraído utilizando o protocolo CTAB padrão [22], com modificações conforme Ferreira & Grattapaglia (1998) [32]. O sequenciamento foi realizado em um sequenciador Illumina GAII a partir de uma biblioteca de fragmentos de DNA total obtida conforme as instruções do fornecedor (www.illumina.com). A amplificação empregando tecnologia Illumina foi gerada por PCR em ponte (*bridge PCR*) de segmentos pareados de leitura (*paired end reads*).

Montagem *de novo* e montagem com referência do genoma cloroplástico de quatro espécies de braquiária

O banco de dados de sequências de *B. ruziziensis, B. brizantha, B. decumbens e B. humidicola* foi usado inicialmente para a montagem com referência do cpDNA destas quatro espécies, usando a sequência completa do cpDNA de *Panicum virgatum* (gb|HQ822121.1|) cultivar *Summer* como comparação. Arquivos de dados no formato FASTQ contendo os segmentos de leitura de sequenciamento foram inicialmente submetidos ao procedimento de correção de sequências curtas do software SOAP*denovo* (Release 1.05), especialmente desenhado para corrigir segmentos do Illumina GA[33]. Em seguida, os segmentos de leitura foram submetidos aos procedimentos de montagem de genomas do software CLC Genomics Workbench 5.1 (CLC Bio, Aarhus, Denmark). Para a montagem *de novo* dos cpDNA das quatro espécies de braquiária, o tamanho da "bolha" de montagem foi automaticamente definida pelo software em 50 pb. Os parâmetros de montagem LF (*Length Fraction*) e Sim (*Similarity*) foram ajustados para 0.5 and 0.8, respectivamente. Os parâmetros MC (*mismatch cost*), DC (*deletion cost*) e IC (*insertion cost*) foram ajustados para 2, 3 e 3, respectivamente. O tamanho do k-mer no montador do CLC Bio foi ajustado para 25 pb e o ponto de corte de cobertura para 30X. Os segmentos de leitura foram montados a partir de bibliotecas de fragmentos (≤200 bp) em contigs usando informações de sobreposição de Kmer (*de Bbruijn graph kmer*). As sobreposições entre os segmentos de leitura foram identificadas pelos grafos De Bruijn [34]. O parâmetro padrão do comprimento de palavra foi ajustado para 25 no CLC Genomics Workbench 5.1.

Anotação de genes do cpDNA e análise de variação de sequência

A anotação dos genes que compõem o cpDNA foi realizada com o programa DOGMA (http://dogma.ccbb.utexas.edu/). A nomenclatura empregada na classificação dos genes do genoma cloroplástico seguiu as regras do Chloroplast Genome Database (http://chloroplast.cbio.psu.edu). DOGMA usa arquivos de entrada no formato FASTA para identificar possíveis genes que codificam proteínas através de procuras BLASTX contra um banco de dados de sequencias de genoma de cloroplasto de diferentes espécies. Para alinhamento e comparação de possíveis regiões gênicas, introns e espaço intergênico dos genomas cpDNA sequenciados foi usado o programa Clustal X 2.0.

Uma montagem "referência" de contigs de *Brachiaria* foi usada para identificar regiões microssatélites (SSR) com o programa PHOBOS (http://www.rub.de/spezzoo/cm/cm_phobos.htm). O número de di-, tri-, and tetra-nucleotídeos de regiões SSR na montagem referência foi computado. Os parâmetros usados pelo programa foram Mismatch score = -5, Gap score = -5 and perfection 100%.

As sequências de cpDNA de braquiária foram submetidas à ferramenta de análise "*SNP detection tool*" do programa CLC Genomics Workbench Version 5.1. Para detecção de SNPs, a sequência cpDNA de *Panicum virgatum* (referência) foi usada como referência, juntamente com os seguintes parâmetros: (a) cobertura mínima de 30x; (b) frequência de variante de 35% para chamada de SNP; (c) alta nota de qualidade (*quality score*) para o SNP e para a região de 11 bases ao redor do sítio SNP.

Identificação, seleção e validação de marcadores indel para diferenciação de espécies de braquiária

Regiões indel do genoma cloroplástico foram identificadas através do alinhamento de montagens *de novo* do cpDNA das quatro espécies de braquiária (*B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola*) usando o programa CLC Genomics Workbench Version 5.1. Os parâmetros usados iniciamente na análise foram: (a) cobertura mínima = 30x; (b) frequência de variante = 20%; (c) número máximo de variações esperadas (ploidia) = 4. Um banco de dados foi criado e utilizado para selecionar um conjunto de indels para validação, levando em consideração parâmetros como posição no genoma cloroplástico, tamanho, cobertura, qualidade da sequência na janela que inclui a indel, e potencial polimorfismo entre as quatro espécies com base no alinhamento dos quatro cpDNA. Os primers para amplificação dos marcadores indel selecionadas foram desenhados com a ferramenta *Primer design* do CLC Workbench

5.1. A validação dos marcadores indel foi realizada através da análise de polimorfismo de DNA por eletroforese em géis de agarose 1% de cada região de inserção/deleção do cpDNA selecionada. Os testes envolveram comparações entre acessos do Banco de Germoplasma de Braquiária, incluindo acessos das espécies *B. ruziziensis, B. decumbens, B. brizantha e B. humidicola.*

Análise filogenética de cpDNA de braquiária

As sequências de cpDNA das quatro espécies de *Brachiaria* (*B. ruziziensis, B. decumbens, B. brizantha e B. humidicola*) foram alinhadas usando o programa ClustalW [35]. Uma análise cladística dos dados de cada loco *barcoding* foi realizada utilizando o critério ML (*Maximum Likelihood*) com o programa MEGA 5 [30], empregando o modelo Tamura-Nei [36]. O suporte para ramificações (*branch support*) dos dendrogramas foi obtido por 100 pseudo-replicações *bootstrap* [31]. Análise filogenética com base em polimorfismo de DNA foi realizada por ML (*Maximum Likelihood*). Como grupo taxonômico externo (*outgroup*) foram utilizadas as sequências completas de outras espécies da família Panicoideae, depositadas no GeneBank, como milho (*Zea mays*; NC_001666.2), sorgo (*Sorghum bicolor*; NC_008602.1), e arroz (*Oryza sativa* sp. *japonica*; NC_001320.1). Os dendrogramas iniciais da procura heurística foram obtidos através da aplicação do método de máxima parcimônia. Uma distribuição Gamma discreta foi usada para modelar a taxa de diferenças evolutivas entre sítios (5 categorias (+G , parâmetro = 0.6909)). O modelo de taxa de variação possibilitou a discriminação de sítios evolutivamente não-variáveis. Todas as posições contendo "gaps" ou dados faltantes foram eliminadas da análise. O dendrograma de máxima parcimônia foi obtido usando o algoritmo *Subtree-Pruning-Regrafting* (SPR) [37]. Os tamanhos das ramificações foram calculados usando o método de caminho médio (*average pathway method*).

Resultados e discussão

Sequenciamento de regiões "*barcoding*" ITS (*internal transcribed spacers 1 e 2, com região central 5.8S*), *trn*H-*psb*A, *rbc*L e *mat*K e análise filogenética de *Brachiaria*

As sequências das regiões "*barcoding*" ITS, *trn*H-*psb*A, *rbc*L e *mat*K das quatro espécies, representadas por seis acessos (*B. ruziziensis* 06 e 10, *B. brizantha* 12 e 14, *B.*

decumbens 18 e *B. humidicola* 19), com 741, 515, 579 e 656 bases de comprimento em cada região, respectivamente, foram alinhadas e analisadas.

As regiões do cpDNA *trn*H-*psb*A, *rbc*L e *mat*K não apresentaram diferenças significativas que dessem suporte a uma análise filogenética. Isto se deve à baixa variabilidade de sequência de DNA detectada nestas regiões nas quatro espécies analisadas. Várias regiões do genoma cloroplástico podem não ter resolução suficiente para a discriminação de espécies dentro do mesmo gênero, especialmente quando a taxa de substituição nucleotídica de cada região é baixa e/ou o período de especiação dos acessos testados é recente [38].

Entre as regiões "*barcoding*" testadas, apenas na região nuclear ITS do rDNA () foi possível detectar variação de sequência entre as espécies de braquiária testadas. Os dados possibilitaram a construção de uma árvore filogenética (Figura 1), onde pode ser observado que *B. ruziziensis, B. brizantha* e *B. decumbens* pertencem ao mesmo grupo monofilético, inclusive com variações de posicionamento no grupo entre os acessos de *B. ruziziensis* 06 e 10, e *B. brizantha* 12 e 14. Observou-se ainda que *B. humidicola* é facilmente separado deste grupo com base nas mutações observadas. A diferenciação entre *B. ruziziensis, B. brizantha* e *B. decumbens* é muito limitada nesta região, referente a um pequeno número de substituições nucleotídicas (entre 2 e 5 bases), ao contrário de *B. humidicola*, onde o número de substituições detectadas foi bem maior (>50 pb). Em gramíneas, a região rDNA tem sido amplamente utilizada em estudos de sistemática molecular, dado que as sequências espaçadoras ITS1 e ITS2 tendem a evoluir mais rapidamente do que a maioria das regiões do cpDNA.

Uma segunda árvore filogenética foi construída com base nas sequências da região nuclear ITS de rDNA de braquiária, desta vez comparando as amostras sequenciadas neste trabalho com um conjunto de acessos de *Brachiaria* e *Urochloa* que possuem sequências depositadas no GenBank. Nesta árvore percebe-se, inicialmente, que o acesso de *B. humidicola* (acesso 19) sequenciado no presente trabalho agrupou com o acesso "controle" de *B. humidicola* depositado no GenBank. Observa-se ainda que estes dois acessos de *B. humidicola* agrupam-se com o acesso de *B. dictyoneura* depositado no GenBank. Na literatura taxonômica de braquiária é comum a confusão entre estas duas espécies. Uma das cultivares comerciais de *B. humidicola*, por exemplo, conhecida como Llanero, tem sido descrita por vezes como *B. humidicola* ou como *B. dictyoneura*. Nesta árvore filogenética os demais acessos testados (*B. ruziziensis 10; B. brizantha 12; B. decumbens 18*) formam um grupo à parte, que inclui

ainda uma amostra de *Urochloa ruziziensis* e *U. brizantha* do GenBank. A similaridade dos gêneros *Brachiaria* e *Urochloa* é, portanto, evidente. Contudo, é importante ressaltar que alguns autores passaram simplesmente a substituir o gênero *Brachiaria* por *Urochloa*, sem que uma análise mais aprofundada com amostras representativas dos dois gêneros tenha sido realizada para diferenciá-los ou unificá-los. Observou-se, por outro lado, que a sequência de uma amostra de *Urochloa decumbens* (ou *Brachiaria decumbens*?) (Figura 2) apresenta-se em outro agrupamento, e várias outras espécies de *Urochloa* estão distribuídas nas diversas ramificações da árvore. Um estudo mais aprofundado da taxonomia *Urochloa/Brachiaria* faz-se, portanto, necessário. De qualquer forma, não resta dúvida que a similaridade de sequências da região nuclear ITS de rDNA das amostras de *B. ruziziensis*, *B. brizantha*, *B. decumbens* usadas no presente trabalho é elevada, indicando uma especiação recente destas espécies.

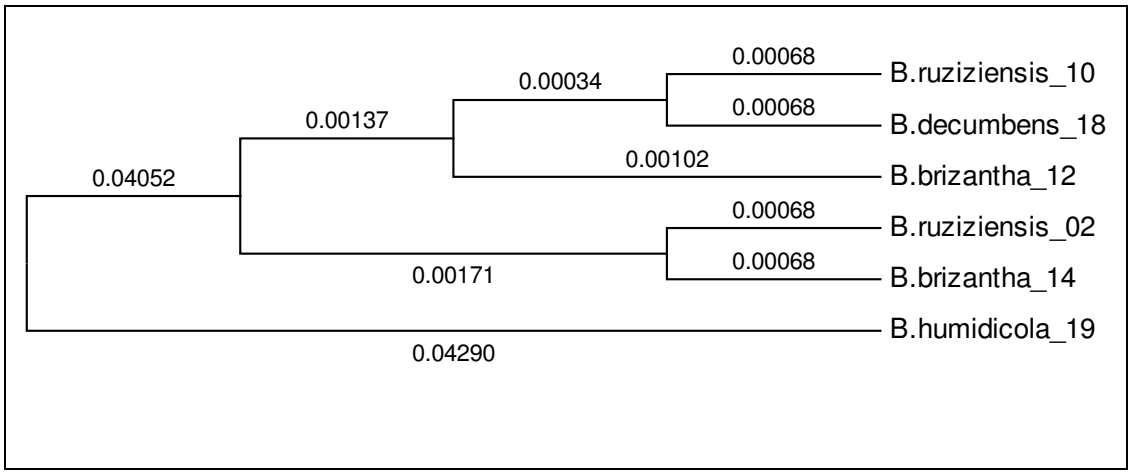


Figura 1 – Árvore filogenética obtida pelo método ML (*Maximum Likelihood)* após alinhamento de sequência de 741 bases da região nuclear ITS de rDNA de seis acessos de *Brachiaria* representando quatro espécies (*B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola*). Log de verossimilhança = -1277.8197.

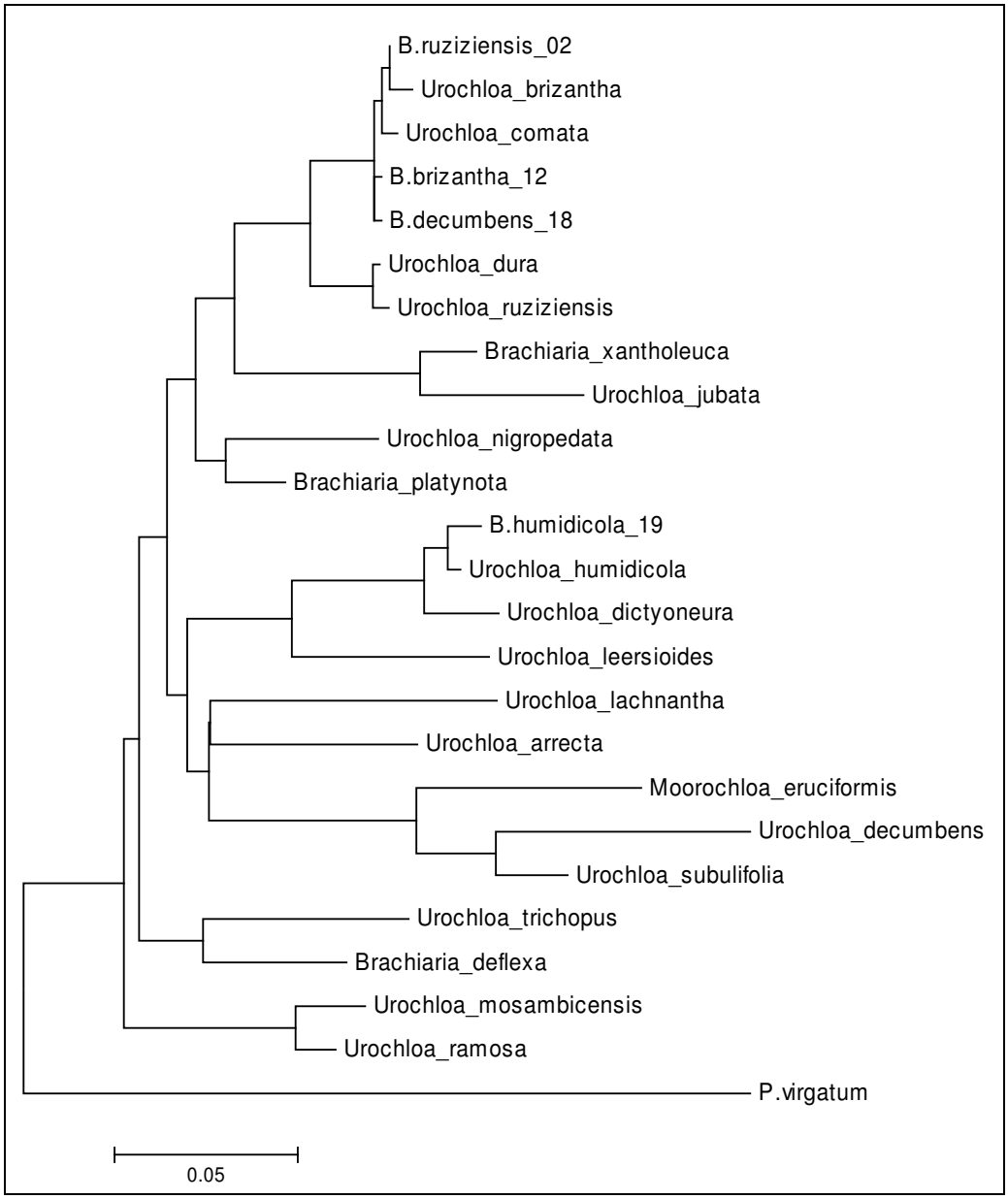


Figura 2 - Árvore filogenética obtida pelo método ML (*Maximum Likelihood)* após alinhamento de sequência de 741 bases da região nuclear ITS de rDNA de acessos de *Brachiaria* representando quatro espécies (*B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola*) e de acessos depositados no GenBank (Gonzalez e Morthon, 2005). Filogenia inferida pelo modelo GTR (*General Time Reversible model*) (GTR). Log de verossimilhança = -2838.4822.

Sequenciamento do genoma de cloroplasto: montagem com genoma de referência

O sequenciamento do genoma cloroplástico geralmente é feito através da extração e separação do cpDNA do genoma nuclear e mitocondrial, seguido por amplificação e purificação para a construção da bibliotecas genômicas [39]. Contudo, neste trabalho optou-se pelo sequenciamento NGS de amostras de DNA total. O sequenciamento NGS apresenta o potencial de sequenciar em conjunto as sequências de genoma nuclear, cloroplástico e mitocondrial, provendo um rendimento de segmentos de leitura dos três genomas. Neste caso, procura-se capitalizar no alto rendimento de segmentos de leitura exclusivos do cpDNA, que potencialmente seriam suficientes para a montagem da sequência do genoma cloroplástico. Como o número de cópias do cDNA é muito elevado em cada célula vegetal em razão da grande quantidade de cloroplastos no citoplasma, o potencial de sucesso do sequenciamento de DNA total para recuperar segmentos de leitura exclusivos do cpDNA é elevado. Genomas cloroplásticos de quatro espécies foram sequenciados desta forma utilizando as seguintes amostras: *B. ruziziensis* FSS-1; *B. brizantha* cv. Marandú, CIAT accession # 6294 (código neste trabalho: acesso 12); *B. decumbens* cv. Basilisky, CIAT accessions # 606 (código neste trabalho: acesso 18); *B. humidicola* cv. Tupi (código neste trabalho: acesso 19).

Foram gerados segmentos de leitura de tamanho médio de 76 pb que, em seguida, foram montados usando o software CLC Genomics Workbench 5.1. O genoma de cloroplasto de *Panicum virgatum* (gb|HQ822121.1|) foi utilizado como referência para a montagem da sequencia completa dos cpDNA das quatro espécies de braquiária. Os resultados de montagem usando um genoma de referência mostraram, inicialmente, que o sequenciamento NGS de DNA total é suficiente para a recuperação de quantidade suficiente de segmentos de leitura exclusivos do cpDNA para a montagem do genoma cloroplástico de cada uma das quatro espécies. Deve ser mencionado que apenas 1/8 da capacidade de corrida NGS da plataforma Illumina GAII gerou segmentos de leitura com cobertura suficiente para a montagem do cpDNA (Tabela 1).

Os quatro genomas de cloroplasto montados possuem uma estrutura circular típica [40], com grande região de cópia única (*Large Single Copy* - LSC) e uma pequena região de cópia única (*Small Single Copy* - SSC), separadas por duas cópias de inversão repetida (*Inverted Repeat* - IR) (Figura 6). O tamanho dos genomas do cloroplasto obtidos variaram entre 138.765 bp em *B. ruziziensis* e 138.976 bp em *B. humidicola*. Entre as milhares de espécies de plantas analisadas, o tamanho do cpDNA varia apenas

de 120 a 210 Kbp. Nenhum dos quatro genomas sequenciados cobriu toda a sequência do genoma referência, o qual apresentou cerca de 800 pb a mais que a sequência do genoma de *Brachiaria ruziziensis,* evidenciando que os genoma de cloroplasto destas espécies de *Brachiaria* parecem ser menores do que a sequência de *Panicum virgatum,* que apresenta 139.619 bases (Tabela 1). Estes dados foram confirmados através da montagem *de novo* dos quatro genomas (veja abaixo).

Embora o número total de segmentos de leitura de *B. ruziziensis* tenha sido cerca de três vezes maior que a quantidade de dados das outras três espécies (*B. brizantha, B. decumbens e B. humidicola*) (Tabela 1), a cobertura e percentual de segmentos de leitura de *B. ruziziensis* mapeados não aumentaram proporcionalmente. Em *B. brizantha,* por exemplo, observou-se um percentual elevado (4%) de segmentos de leitura mapeados no genoma cloroplástico de *P. virgatum*, comparado aos cerca de 1-2% de segmentos de leitura mapeados nas outras três espécies, incluindo *B. ruziziensis* (Tabela 1). O melhor aproveitamento dos segmentos de leitura ocorrido em *B. brizantha* resultou em uma cobertura média do cpDNA (2.791x) superior à observada em *B. ruziziensis* (2.011x), apesar da quantidade inicial de segmentos de leitura *de B. ruziziensis* ter sido três vezes maior. Os dados sugerem uma potencial maior proporção de cpDNA nas amostras de DNA total usadas no sequenciamento NGS de *B. brizantha* do que das outras espécies.

Observou-se que a cobertura média dos contigs de referência foi muito alta (Tabela 1), excedendo 1.000 X. Isto não parece estar relacionado com o número de segmentos de leitura inicial obtido para cada espécie, visto que *B. brizantha* teve melhor cobertura média do cpDNA do que as demais espécies.

Chama a atenção o fato do tamanho final das sequências de cpDNA montados para *B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola* serem tão próximas (138.765, 138.945, 138.940, 138.976). A diferença foi de apenas 5 pb entre *B. brizantha* e *B. decumbens*, e de até 208 bp entre *B. humidicola* e *B. ruziziensis*.

Tabela 1- Métricas do sequenciamento e montagem do genoma de quatro espécies de *Brachiaria* usando *P. virgatum* (cp) como genoma de referência.

Espécie	Segmentos de leitura Total (bp)	Segmentos de leitura mapeados no cpDNA de *P. virgatum* (bp)	Segmentos de leitura não mapeados no cpDNA de *P. virgatum* (bp)	% de segmentos de leitura mapeados	Tamanho do cpDNA montado (bp)	Cobertura média
B. ruziziensis	20.211.010.448	279.025.488	19.931.984.960	1%	138.765	2.011
B.brizantha	8.643.705.720	387.850.876	8.255.854.844	4%	138.945	2.791
B.humidicola	8.476.910.040	183.602.548	8.293.307.492	2%	138.976	1.321
B. decumbens	9.018.811.776	168.717.644	8.850.094.132	2%	138.940	1.214
P. virgatum (reference)					139.619	

Sequenciamento do genoma de cloroplasto: montagem *de novo*

Os dados de sequenciamento NGS foram utilizados também em procedimentos de montagem *de novo* do genoma cloroplástico das quatro espécies de *Brachiaria* para fins de comparação com os resultados obtidos com a montagem com referência utilizando a sequência de cpDNA de *P. virgatum*. Observou-se, novamente, que *B. brizantha* apresentou melhores parâmetros de montagem do que *B. humidicola* e *B. decumbens*. O N50 do tamanho dos contigs resultantes do sequenciamento *de novo* do cpDNA de *B. ruziziensis* (1.704) foi cerca de três vezes maior do que o obtido para os três demais genomas, que tiveram um N50 variando de 485 a 505 bases. Contudo, os maiores contigs montados não foram de *B. ruziziensis,* mas sim os de *B. brizantha* e *B. humidicola*.

Tabela 2 – Parâmetros de sequenciamento e montagem *de novo* do genoma de quatro espécies de *Brachiaria* .

Espécie	Segmentos de leitura (pb)	N50	Tamanho mínimo (pb)	Tamanho máximo (pb)	Tamanho médio (bp)	Número de contigs	Total (bp)
B. ruziziensis	20.211.010.488	1.704	200	57.461	754	382.380	288.171.438
B.brizantha	8.643.705.720	505	200	86.745	464	380.899	176.721.916
B.humidicola	8.476.910.040	485	200	80.478	448	476.044	213.428.620
B. decumbens	9.018.811.776	491	200	38.791	455	408.401	185.680.658

A montagem de pequenos segmentos de leitura de DNA é uma estratégia desafiadora, mas que se apresenta rápida e eficiente na montagem de cpDNA. O emprego de genoma cloroplástico como referência certamente facilita o processo de montagem, conforme verificado neste e em outros estudos [41, 42]. A montagem *de novo* pode ser também considerada muito eficiente, dado o nível de recuperação da sequência de cpDNA que se obtém. A cobertura linear alcançada neste trabalho pelo somatório das sequências montadas (*scaffolds*) que alinharam com e-value = 0 variou entre 92,89 a 99,45%, considerando a duplicidade da região IR, que na montagem *de novo* mapeou em uma única região. A Tabela 3 apresenta os *scaffolds* mais representativos que alinharam sem sobreposição.

Tabela 3 – Número do *scaffold* e tamanho em número de bases da montagem *de novo* do cpDNA de quatro espécies de *Brachiaria*, que alinharam com o cpDNA de referência de *P. virgatum* (e-value = 0). Os *scaffolds* grifados em negrito correspondem às duas *inverted repeats* (IR) combinadas e foram considerado em dobro para avaliação da cobertura linear.

Identificação do scaffold RUZI	Tamanho (pb)	*Identificação do scaffold HUMI*	Tamanho (pb)	*Identificação do scaffold BRIZ*	Tamanho (pb)	*Identificação do scaffold DEC*	Tamanho (pb)
303	12642	275	12548	304	12668	94	12626
211	57438	9	80455	196	36994	63	33833
10	23716	9	.	46	23519	1015	2217
19	**20396 x 2**	**142**	**22605 x2**	**30**	**22642 x 2**	**106**	**20444 x 2**
623	2145	-		14	15700	356	10664
-	-	-	-	-	-	3	27337
-	-	-	-	-	-	1410	1491
Total pb	136733		138213		134165		129056
Cobertura	98,5%		99,45%		96,50%		92,89%

RUZ= B. ruziziensis; HUM= B. humidicola; BRIZ= B. brizantha; DEC= B. decumbens

Desenvolvimento e validação de marcadores Indel para uso como ferramenta de diferenciação de espécies de *Brachiaria*

Alterações microestruturais, tais como pequenas inserções e deleções (indels) do cpDNA, ou até mesmo inversões, podem ser extremamente úteis para resolver relações filogenéticas entre acessos de um mesmo gênero [43], para inferir as relações de vínculo genético entre acessos mais relacionados [44], ou para serem usadas para a rápida discriminação de espécies em programas de conservação de germoplasma. Isto pode ser particularmente importante entre acessos de grupos morfologicamente muito semelhantes, onde há grande dificuldade de separação de espécies pela ausência de claros descritores morfológicos, como ocorre nas braquiárias. A existência sistemas de análise de alterações microestruturais do genoma possibilita, por exemplo, a coleta de acessos a qualquer momento e posterior classificação taxonômica dos mesmos através de ensaios laboratoriais rápidos e eficientes.

A identificação de regiões indel entre as quatro espécies de *Brachiaria* foi realizada pelo alinhamento da montagem d*e novo* do cpDNA de cada espécie. Inicialmente, cada cpDNA foi mapeado no genoma referência de *Panicum virgatum* com uso do software Blast. Para isto foram

separadas as sequências mais longas e de maior qualidade dos quatro genomas montados *de novo*, e que melhor mapearam no genoma de referência (*e-value* = -10). Tomando-se, por exemplo, os scaffolds da montagem *de novo* do cpDNA de *B. ruziziensis* selecionados após o alinhamento, observou-se ao menos um outro *scaffold* montado *de novo* das outras três espécies (Tabela 4). Estes scaffolds foram alinhados para que pudesse ser feita a detecção de indels no cpDNA das quatro espécies. A única exceção foi um pequeno *scaffold* de 2.145 pb, localizado na região de IR (*scaffold 2*) de *B. ruziziensis*, para o qual não foi encontrada uma região comparativa de cpDNA nas outras três espécies (número 623 na Tabela 4).

Observou-se que os dois *scaffolds* da região IR do cpDNA de *B. ruziziensis* (*scaffolds* 19 e 623) somaram 22.541 pares de bases. Isto é consistente com os tamanhos médios, entre 20 e 30 Kb, de um único segmento de IR de cloroplasto de angiospermas [40]. Assim, presume-se que a sequência final IR montada *de novo* seja uma combinação dos dois segmentos IR do cloroplasto de *B. ruziziensis*. Em outras palavras, a montagem *de novo* não possibilitou a discriminação das duas regiões *inverted repeats* (IR), visto que os segmentos de leitura das duas IRs (*forward* e *reverse*) foram mapeados na mesma região. Isto foi comprovado ao se mapear os segmentos de leitura de *B. ruziziensis* nas duas *inverted repeats* do genoma cloroplástico de referência (*P. virgatum*), que resultou nos seguintes valores: IR1 (cobertura= 2.409; tamanho=24.448 pb); IR2 (cobertura=2.467; tamanho=23.983). Portanto, os valores de cobertura na região IR do cpDNA de *B. ruziziensis* (5.050 e 4.167) são o dobro da cobertura observada no mapeamento nas regiões IR1 e IR2 do genoma cpDNA referência (2.409 e 2.467).

Tabela 4 – Cobertura observada e tamanho (pb) de *scaffolds* obtidos na montagem *de novo* dos quatro cpDNA das espécies *B. ruziziensis, B. humidicola, B. brizantha* e *B. decumbens*. Os números que identificam os *scaffolds* correspondentes de cada espécie para as regiões IR, LSC e SSC do genoma do cloroplasto são apresentados.

			identificação do scaffold			
Região do cpDNA	cobertura	Tamanho (pb)	RUZ	HUM	BRIZ	DEC
SSC	1.846	57.461	303	275	304	94
LSC (scaffold 1)	1.470	12.642	211	9	196	63
LSC (scaffold 2)	1.637	23.716	10	9	46	1015
IR (scaffold 1)	5.050	20.396	19	142	30	106
IR (scaffold 2)	4.167	2.145	623	-	-	-
P. virgatum IR1	2.409	24.448		-	-	-
P. virgatum IR2	2.467	23.983		-	-	-

RUZ= B. ruziziensis; HUM= B. humidicola; BRIZ= B. brizantha; DEC= B. decumbens

O alinhamento da montagem *de novo* do cpDNA das quatro espécies possibilitou a identificação de um grande número de indels. Estes eventos de inserção/deleção são em geral atribuídos à repetição perfeita ou quase perfeita de uma seqüência adjacente, provavelmente causada por escorregamento na replicação do DNA [45]. Estima-se que o número de indels curtas (1-10 pb) em gramíneas representam mais de 90% do total de indels detectadas e que as espécies mais estreitamente relacionados tendem a ter uma maior proporção de indels curtos [46]. Com base nisto, e tendo como ponto de partida os *scaffolds* referenciados na Tabela 4, foram selecionados 18 indels que apresentam polimorfismo de inserção/deleção *in silico* e permitem distinguir as quatro espécies de *Brachiaria* que tiveram o cpDNA sequenciado (*B. ruziziensis, B. brizantha, B. decumbens e B. humidicola*). O propósito foi identificar os sítios mais polimórficos de inserção/deleção para testá-los na discriminação das espécies em testes de laboratório. Um conjunto de *primers* flaqueando o sítio de inserção/deleção destas 18 regiões (indels) foi, então, desenhado para testes laboratoriais. A Tabela 5 mostra o conjunto de marcadores indel desenvolvidos para discriminação de espécies de *Brachiaria*. A Figura 3 apresenta o resultado de discriminação de diferentes espécies de *Brachiaria* com alguns dos marcadores indels desenvolvidos no presente trabalho. Observa-se, neste exemplo, que com os marcadores 66584, 107669 e com o multiplex dos marcadores 93252 e 107669 pode-se separar facilmente os acessos de *B. ruziziensis* dos acessos de *B. brizantha, B. decumbens e B. humidicola*. Da mesma forma, acessos de *B. humidicola* são separados dos demais acessos com os marcadores 93252 e combinação em multiplex dos marcadores 93252 e 107669. Note-se, contudo, que o tamanho do indel que separa *B. brizantha* de *B. decumbens* nestes marcadores é, em geral, muito pequeno, o que não permite a separação de produtos de PCR em gel de agarose. Contudo, acessos *de B. brizantha* podem ser separados de acessos de *B. decumbens* com o emprego de eletroforese em gel de poliacrilamida (Figura 4).

Table 5 – Indel "primers" para separação de acessos de espécies de *Brachiaria* desenvolvidos a partir da montagem *de novo* do genoma de cloroplasto. Os números de referência indicam a posição no genoma de cloroplasto de *P. virgatum*, números com dupla referência referem-se a posições em regiões IR. Marcadores entre as posições 107669 e 114885 estão localizados na SSC e após 81.616 na região LSC. .

					Primers		Tamanho Esperado do Fragmento (bp)			
Ref.	Scaffold #	Tamanho (bp)	tipo	Identifica	FWD	REW	*RUZI*	*DEC*	*BRI*	*HUM*
66584	10	20	ins	*B. ruziziensis*	AAGAAGTTCTTACTCTTTCTGT	ACATACGACTCATAATGAA	**105**	**83**	74	74
72645	46	6	ins	*B. brizantha*	GAAAGAGAAAAAAGTTGTC	AGAGTGGATCAAGAAAAAA	153	153	**161**	153
72956	10	5	ins	*B. ruziziensis*	TCATCTGTCTTTCTTTCC	CTATCAGAAAACCACTAT	**175**	130	130	130
74248	10	6	ins	*B. ruziziensis*	CGATGCAAAGAAAATGAATG	CGTAAGATCCCATAGAGT	**119**	113	113	113
75494	10	12	ins	*B. ruziziensis*	AGTTCTCGCTTTAAATCC	CCCTAGATACCTAAAATC	**193**	150	148	145
79281	10	5	ins	*B. ruziziensis*	GCCCGCGAAATCCTTATT	CAAAACTGGACATGAGAG	**162**	157	157	157
81154	9	51	ins	*B. humidícola*	TGAAGTCAGTAGGAGT	GGAATCGAAATCTTGG	153	153	153	**203**
81616	10	6	ins	*B. ruziziensis*	AAAGATTCAGAATAAACAAA	GAAGAAGAACGGGCTAAGGAAA	**149**	143	143	**135**
107669	303	19	ins	*B. ruziziensis*	CGAGCATCCAAAACCAAAA	ATGGATAACGGAGGGATT	**224**	**203**	**213**	**199**
113003	275	50	ins	*B. humidícola*	CAAGGAAGGAAAAAGATA	AGTAAACTAGACGAAGAA	177	176	176	**126**
114885	303	5	del	*B. ruziziensis(2)*	TTTCTAATCCCTCACTAAC	GTAAACATAAGCAGTGTAA	**177**	182	182	182
119374/101302	106	6	ins	*B. decumbens*	CTTCTTCTCCTCAGCCATT	CATCACATCCCCTCTCTC	**109**	104	104	**109**
103778/117488	19	6	ins	*B. humidícola*	ATTGGATTTGGATAGAAGGGTA	GCAATAAAAAAATCAGCAAAATTC	95	93	95	**88**
86220/135017	19	6	ins	*B. ruziziensis*	GTTAGATAGGAACAGCTTTG	TTTATGAACGGGAATGGG	**121**	116	116	116
87460/133763	19	5	ins	*B. ruziziensis*	TAAGTAGCGATCAAGGAA	GCTCAAAGAACGAATAAA	**123**	118	118	118
93252/127974	142	22	ins	*B. humidícola*	CACGGAAGAAAGAACTCA	CGGGGAAAGTATACAGAAAA	157	157	157	**180**

RUZ= B. ruziziensis; HUM= B. humidicola; BRIZ= B. brizantha; DEC= B. decumbens

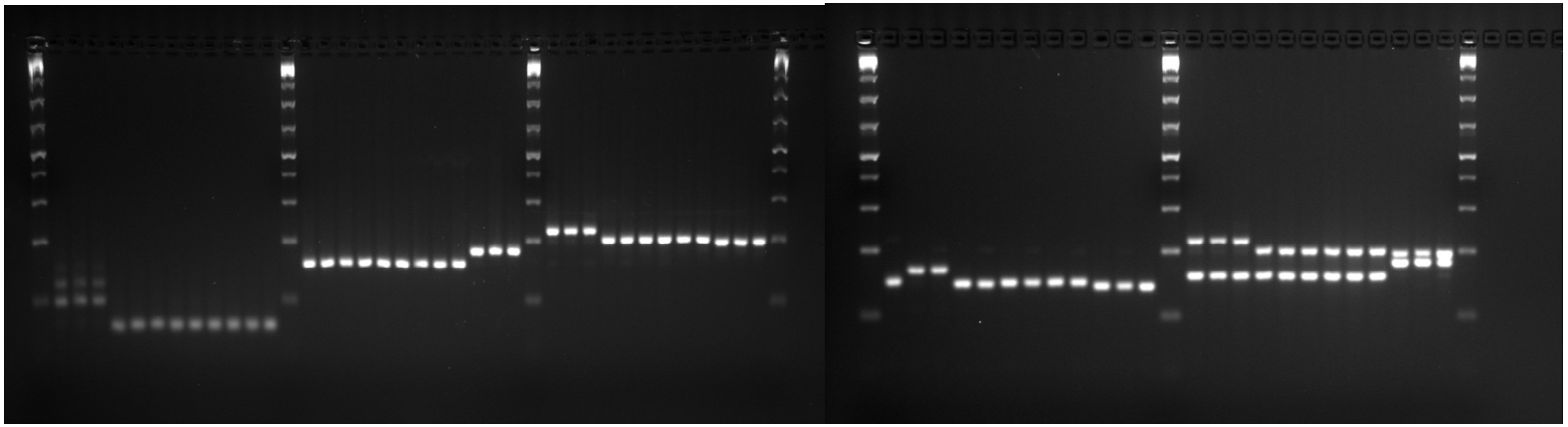


Figura 3 – Discriminação de acessos de quatro espécies de *Brachiaria* com marcadores indel selecionados no cpDNA. Polimorfismo de DNA de amostras de diferentes acessos do Banco de Germoplasma submetidas a eletroforese em gel de agarose 1%. Trêsmarcadores são apresentados (da esquerda para a direita): 66584, 93252, 107669, além da combinação em multiplex dos marcadores 93252 e 107669. As amostras das diferentes espécies são apresentadas na seguinte ordem para cada marcador ou multiplex (da esquerda para a direita):*B. ruziziensis* (Kennedy, BRA-5541-00, BRA-5550-00), *B. brizantha* (Marandu, BRA-000591, BRA-001384), *B. decumbens* (Basiliski, BRA-000116, BRA-001058), *B. humidicola* (Tupi, BRA-001929, BRA-001937). Os marcadores são separados pela escada alélica (ladder) 50 pb (Promega).

.

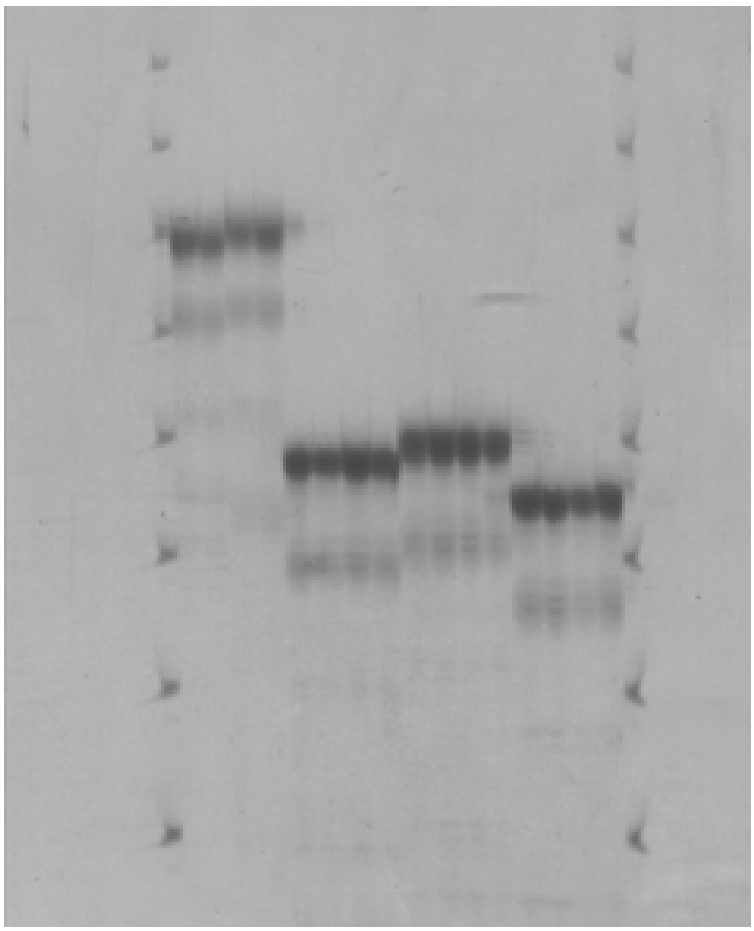


Figura 4. Discriminação de acessos de quatro espécies de *Brachiaria* com marcador indel selecionado no cpDNA. Polimorfismo de DNA no loco indel RUBRIZ (ref. 75494) entre amostras de diferentes acessos do Banco de Germoplasmasobmetidas a eletroforese em gel de poliacrilamida. As amostras das diferentes espécies são apresentadas na seguinte ordem: *B. ruziziensis* (Kennedy, BRA-5541-00), *B. brizantha* (Marandu, BRA-000591), *B. decumbens* (Basiliski, BRA-000116), *B. humidicola* (Tupi, BRA-001929). Os marcadores são separados pela escada alélica (ladder) 50 pb (Promega). As amostras foram repetidas lado a lado, em testes de prova e contra-prova.

Anotação de genes do cpDNA de *Brachiaria*

O genoma do cloroplasto das quatro espécies de *Brachiaria* contém 118 genes únicos, dos quais 18 são duplicados nas regiões invertidas IRs, perfazendo um total de 136 genes de função conhecida (Figura 5). Além disso, existem nove ORFs e três pseudogenes. Na Figura 5 é apresentada a estrutura e o mapeamento dos genes identificados no cpDNA de *B. ruziziensis*, incluindo a identificação das regiões IR, LSC e SSC. Figuras semelhantes foram obtidas para *B. brizantha, B. decumbens e B. humidicola* (não apresentadas). O número de genes e ordem dos mesmos na estrutura linear do cromossomo circular são idênticos aos descritos em outros genomas de cloroplastos de gramíneas, tal como relatado por Bortiri et a. (2008) [39] em estudos sobre *Brachypodium*, milho e trigo. A relação dos genes identificados é apresentada no Anexo 4.

Microssatélites no cpDNA

A variação de sequências de microssatélites (*Single Sequence Repeat* - SSR) do genoma de cloroplastos pode ser também utilizada como marcador molecular em análise genética. Os SSRs geralmente têm uma maior taxa de mutação em comparação com outras regiões neutras de DNA e em genomas de cloroplasto, comumente, apresentam variação intraespecífica em número de repetição do motivo. A identificação e o estudo de sequências SSR no cpDNA de braquiária pode ser potencialmente importante para o desenvolvimento de ferramentas para apoio a programas de uso e conservação de recursos genéticos, desde que seja demonstrado a sua abundância, facilidade de detecção e grau de polimorfismo de DNA.

Como outros marcadores de cloroplasto de herança uniparental, os marcadores SSR de cloroplasto (cpSSR), também podem ser usados em análise de estrutura populacional de plantas, diversidade genética, diferenciação populacional e análise de maternidade (herança materna). As variações inter e intra-específicas reveladas por cpSSR foram estudadas em populações de plantas, incluindo muitas espécies de Poaceae [47].

Tabela 6 - Número de SSRs perfeitos com variações de di, tri e tetra nucleotídeos encontrados nas seqüências montadas de cpDNA de *Brachiaria*, tendo o genoma de cloroplasto de *P. virgatum* como referência. Os motivos de sequência microssatélites mais abundantes foram anotados e quantificados.

	Total	Di	Tri	Tetra	(AAG)	(AT)	(AAT)	(AG)	(AAAG)
B. decumbens	1452	302	553	597	165	144	135	117	104
B. brizantha	1461	303	556	602	168	144	135	117	105
B. ruziziensis	1440	300	544	596	165	144	124	115	109
B. humidicola	1471	305	562	604	168	147	139	119	105
P. virgatum	1487	292	565	630	165	144	135	117	104

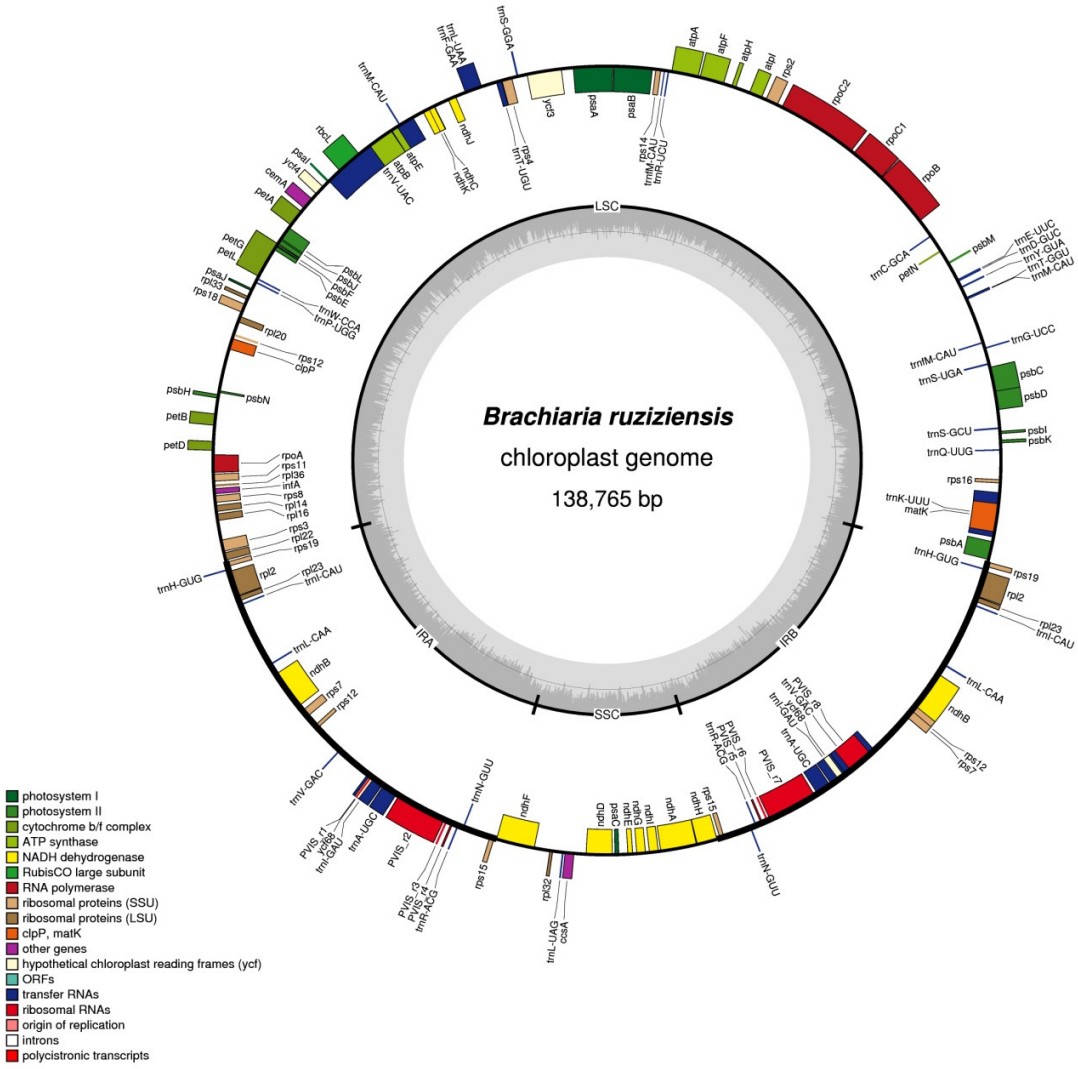


Figura 5. Mapa genético do genoma do cloroplasto de *Brachiaria ruzizensis*. O mapa inclui as repetições invertidas, IRa e IRb, regiões de cópia única pequena (SSC) e grande (LSC). Genes identificados no interior do mapa são transcritos no sentido horário, enquanto que os genes do exterior do mapa são transcritos em ordem inversa.

Os resultados da análise comparativa do número de polimorfismos de cpSSRs de di, tri e tetranucleotídeos entre as quatro espécies de braquiária possibilitaram a identificação de 1440 a 1471 cpSSRs com motivos de di, tri e tetranucleótideos, que foram identificadas e anotadas nos genomas montados usando o cpDNA de *P. virgatum* como referência. As repetições de tetranucleotídeos foram as mais abundantes (~40%), o que é consistente com os observado em outros genomas de gramíneas, como *Panicum virgatum* [46, 48]. As repetições predominantes foram AAG (12%), AT (10%), AAT (10%), AG (8%) e AAAG (7%), nos quatro genomas (Tabela 6).

Os resultados de testes *in silico* indicaram que o polimorfismo de cpSSR é limitado entre os genomas cloroplásticos das quatro espécies *Brachiaria.* Na maior parte dos microssatélites analisados não foi detectado polimorfismo de repetição de sequência. O limitado polimorfismo de cpSSR entre as quatro espécies dificulta a seleção de marcadores informativos para análise genética de braquiária. Isto provavelmente se deve ao fato de que grandes repetições são suprimidas ou eliminadas seletivamente do DNA de cloroplasto devido à sua capacidade de desestabilizar a estrutura deste genoma [49, 50]. O número de SSR perfeitos anotados foi maior para os maiores motivo (tetra>tri>di).

SNPs e Indels no cpDNA

Neste estudo foram detectados SNPs no cpDNA de braquiária baseados em comparações par-a-par entre as quatro espécies consideradas. Verificou-se um número de SNPs em comparações par-a-par variando de 50 SNPs, existentes entre as sequências de cloroplasto de *B. brizantha* e *B. decumbens*, a 1018 SNPs, identificados na comparação entre as sequencias de *B. decumbens* e *B. humicola*. A relação entre o número de variações SNPs apoia os resultados da análise de filogenia registrados anteriormente, visto que a menor quantidade de SNPs computados nas comparações par-a-par foi encontrado entre as espécies filogeneticamente mais próximas. Estes resultados mostraram, por exemplo, que a *B. ruziziensis* é mais próxima geneticamente de *B. brizantha* do que de *B. humidicola*. Mostraram ainda que a semelhança entre *B. brizantha e B. decumbens* é alta, o que foi caracterizado pelo baixo número de SNPs encontrados na comparação entre estas duas espécies (Tabela 6).

As comparações par-a-par das sequências de cpDNA que foram montadas permitiram

ainda identificar indels com comprimento variando entre 1 e 8 pares de bases entre as espécies estudadas. Verificou-se, por exemplo, que o menor número de indels (77) foi computado entre *B. brizantha* e *B.decumbens*. Já o maior número de indels (301) foi identificado *entre B. decumbens* e *B. humidicola*. A Tabela 6 apresenta a quantidade de indels identificadas nas comparações par-a-par entre as quatro espécies de *Brachiaria*. Como esperado, as indels de um único nucleotídeo (1 base) foram as mais comuns, representando pelo menos 40% do total. As espécies mais próximas apresentam um menor número de indels entre si (ex. *B. brizantha* e *B. decumbens*). No entanto, o número de indels identificadas não diminuiu com o aumento do tamanho da inserção/deleção. Por exemplo, as indels com 5 bases são mais abundantes entre as espécies (~12%) em comparação com os indels de 3 ou 4 bases comprimento. Fenômeno similar foi observado em outras espécies de gramíneas [46]. Estimativas semelhantes foram relatadas em estudos com cana, arroz e milho [51]. Foram identificados apenas 177 indels entre *B. brizantha* e *B. decumbens*, muito menos do que aquelas encontrados nas outras comparações pareadas, novamente indicando a proximidade entre estas duas espécies. Além disso, observou-se a elevada correlação (0,856512) entre o número de indels e o número de SNPs encontrados nas comparações interespecíficas (Tabela 7).

Tabela 7 - Número de indels e SNPs entre seqüências de cpDNA de quatro espécies de *Brachiaria* comparadas par-a-par. *B. humidicola* (Hum), *B. ruziziensis* (Ruzi), *B. decumbens* (Dec) e *B. brizantha* (Briz). A correlação entre Indel e SNPs é 0,856512.

	Comprimento da inserção/deleção (indel) (bp)								Total	
Comparação par-a-par	1	2	3	4	5	6	7	8	Indel	SNP
Hum x Ruzi	151	40	25	28	34	10	0	1	289	896
Briz x Ruzi	112	34	21	18	26	10	1	0	222	359
Dec x Ruzi	112	37	25	21	30	11	1	0	237	525
Briz x Hum	162	44	25	17	25	11	1	0	285	398
Briz x Dec	72	31	21	17	22	11	3	0	177	50
Dec x Hum	168	45	23	22	29	13	0	1	301	1018

Filogenia e estimativa de tempo de divergência

Para a avaliação das relações filogenéticas entre as quatro espécies de *Brachiaria* foi realizada uma análise filogenética pelo método de máxima verossimilhança (ML) utilizando as sequências obtidas para cada montagem *de novo* do cpDNA. Os resultados indicam uma mesma topologia de árvore filogenética para os três conjuntos das regiões LSC, SSC e IR do cpDNA (Figura 6). Árvore inicial (s) para a busca heurística foi obtida automaticamente através da aplicação de

Neighbor-Join e algoritmos BioNJ a uma matriz de distâncias estimadas entre pares usando a abordagem de probabilidade de composição máxima (MCL) e, em seguida, selecionada a topologia com o valor de verossimilhança superior. As árvores estãp desenhadas em escala (Figura 6), com comprimentos dos ramos medidos no número de substituições por sítio (acima dos braços). A análise envolveu quatro sequências de nucleotídeos. Posições do códon incluídas foram 1ª, 2ª, 3ª + não-codificante. Todas as posições que contêm lacunas e dados faltantes foram eliminadas. Um total de 12.494, 8539 e 20350 posições foram consideradas no conjunto de final de dados. Como esperado, em consequência das similaridades de sequencia de cpDNA já verificadas nas análises de rDNA, SNPs e indels apresentadas anteriormente, as espécies *B. decumbens* e *B. brizantha* apresentam a maior proximidade.

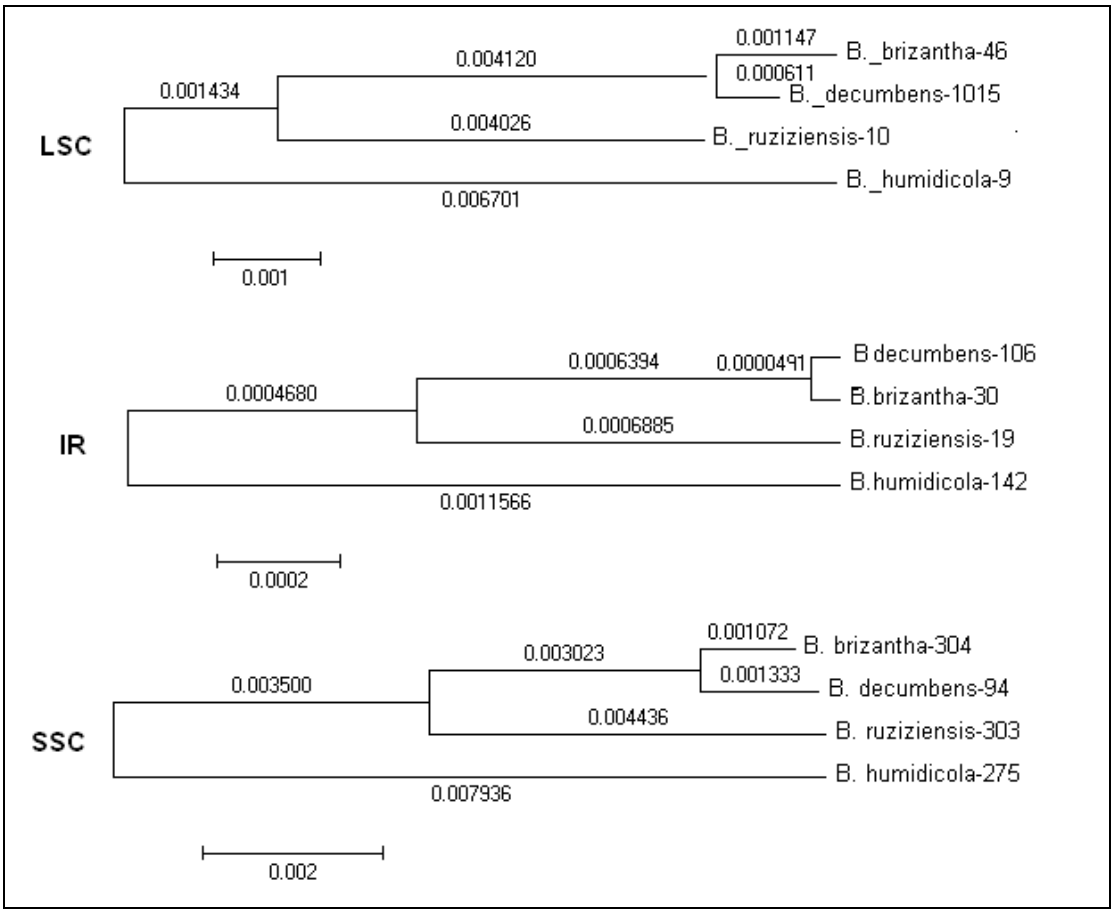


Figura 6.Dendrogramas baseados na sequência de DNA de diferentes regiões do cpDNA (LSC, IR e SSC) pelo método Maximum Likelihood baseado no modelo Tamura-Nei. As árvores com a maior verossimilhança (LSC =-18.409,3501, IR (combinado) = -12545.7330 e SSC = -28603.4495 são apresentadas. Números ao lado dos nomes das espécies representam o nímero do contig na montagem "de novo".

As relações filogenéticas das quatro espécies de *Brachiaria* foram ainda exploradas por uma análise comparativa das quatro sequências completas do cpDNA (usando referência), juntamente com sequências completas de *Panicum virgatum, Oryza sativa, Zea mays e Sorghum bicolor* do Genbank. Estas sequências foram alinhadas e usadas para reconstruir as topologias de árvores filogenéticas por máxima parcimônia e pelo método de máxima verossimilhança (ML). A análise envolveu oito sequências de cpDNA e todas as posições que continham lacunas ou dados faltantes foram eliminadas.

A árvore com a maior verossimilhança (-243.395,9130) foi concebida com modelo de variação evolutiva invariável (1,96807 x 10^{-9}) e 1000 réplicas de "bootstrap", usando um total de 128.636 posições no conjunto final de dados. O relógio molecular foi calibrado usando o ponto de divergência de *Oryza sativa* e *Zea mays*, com ocorrência estimada há 65 milhões de anos (MYA) [52]. A substituição modelo foi definida com um GTR + G + I, determinado como o melhor ajuste modelo através de testes de razão de verossimilhança hierárquicos.

Todas as árvores filogenéticas construídas tiveram a mesma topologia e indicam que *B. decumbens* e *B. brizantha* são derivadas de um ancestral comum com *B. ruziziensis* (Figuras 7 e 8). Indicam ainda que *B.humidicola* é a espécie mais distante nas comparações de polimorfismo de cpDNA com *B. ruziziensis, B. decumbens e B. brizantha*. Para efeito de estimativa de tempo separação destas espécies, a divergência de sequência de cpDNA aponta para uma taxa evolutiva total para cloroplasto = 1,96807 X 10^{-9}. Este valor está de acordo com os limites atribuídos à taxa de substituição média e nucleotídeos de genes de cloroplastos, estimado em cerca de 1,1 a 2,9 × 10^{-9} substituições sinônimas por sítio nos estudos da história evolutiva das plantas realizados por Muse (2000) e Jakobsson et al. (2007) [53, 54]. Com base nestes parâmetros, estima-se que o tempo de divergência entre *B. decumbens* e *B. brizantha* seja 2,5 MYA, e que estas duas espécies tenham se separado do ancestral que originou *B. ruziziensis* há 14 MYA (Figuras 7 e 8). Isto provoca a hipótese de que *B. decumbens* e *B. brizantha* tenham surgido do ancestral de *B. ruziziensis* por evento(s) de poliploidização, que isolou reprodutivamente estas espécies.

As estimativas de divergência evolutiva entre as sequências, na qual baseiam-se a construção das árvores filogenéticas, possuem baixo valor de erro padrão (Tabela 8), indicando que a topologia mais parcimoniosa é a que foi apresentada. A análise Bayesiana e a de máxima verossimilhança (ML) produziram árvores semelhantes em cada partição do genoma de cloroplasto e as árvores filogenéticas das quatro sequências foram congruentes. A árvore também

se mostrou congruente com a posição de *Zea, Sorghum, Panicum e Oryza* [42], bem como com o agrupamento de espécies da subfamília Panicoidea obedecendo a distribuição no clado PACC (Panicoideae, Arundinoideae, Chloridoideae e Centothecoideae) [55], revisto pelo Grass Phylogeny Working Group [56].

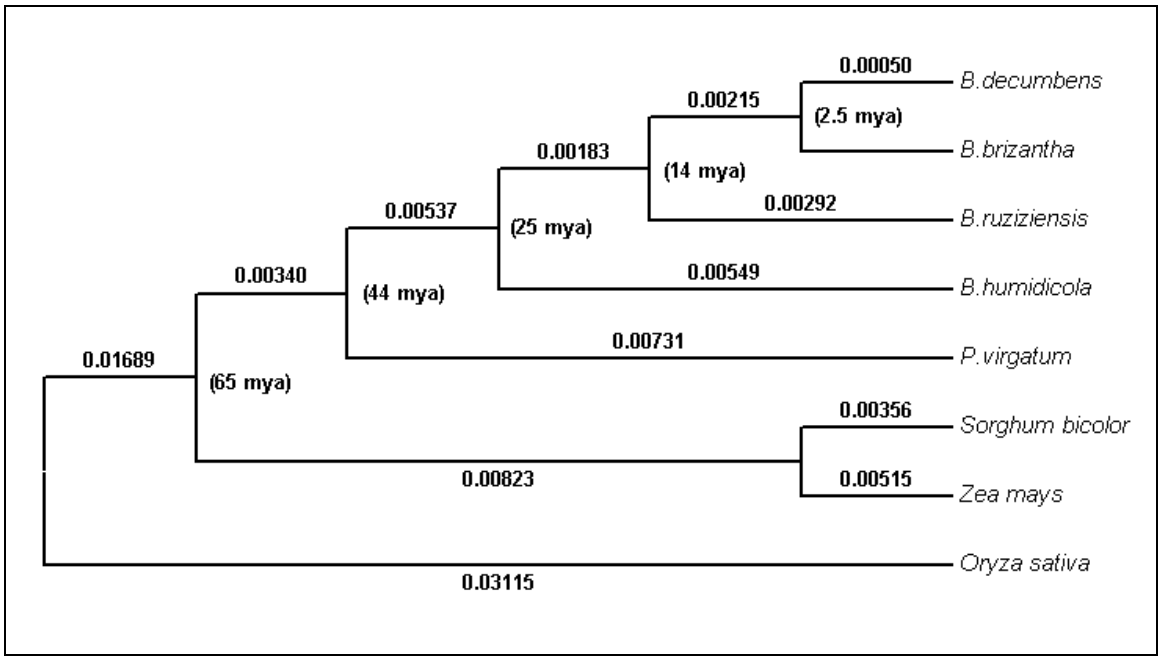


Figura 7. A árvore filogenética construída por ML apresentando a maior verossimilhança = -243.395,9130. O modelo de variação da taxa foi evolutivamente invariável. Um total de 128.636 posições nucleotídicas foram consideradas no conjunto de dados utilizado na análise. O relógio molecular foi calibrado usando um ponto de divergência de *Oryza sativa* e *Zea mays* com ocorrência há 65 milhões de anos (MYA). Taxa Evolutiva = 1,96807 X 10^{-9}. O número de repetições no teste de "bootstrap" foi 1000. Números entre parêntesis correspondem à estimativa de tempo de divergência com comprimentos medidos no número de substituições por sítio (acima dos ramos)

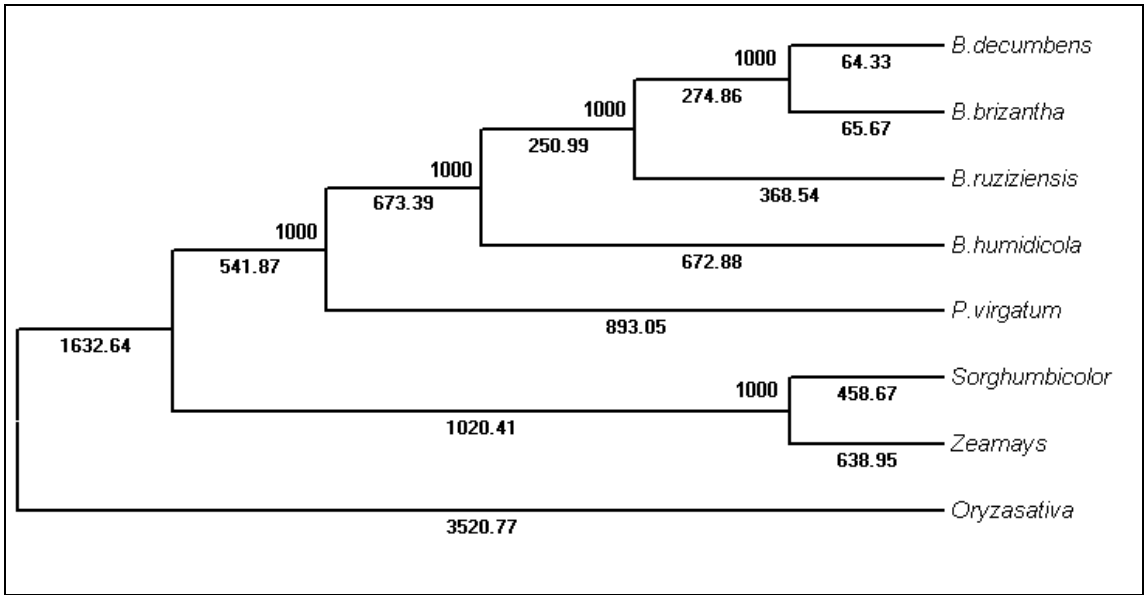


Figura 8 - A árvore filogenética construída por máxima parcimônia. A árvore mais parcimoniosa com comprimento = 11077 é apresentada. O número de árvores idênticas computadas por teste de "bootstrap" (1000 réplicas) são mostradas ao lado dos ramos. Os comprimentos dos ramos foram calculados usando o método da média de percurso e está em unidades de número de mudanças ao longo de toda a sequência. Todas as posições que contêm lacunas e dados faltantes foram eliminadas da análise. Um total de 128.636 posições nucleotídicas foi considerado no conjunto de dados utilizado na análise.

Tabela 8. Estimativas de divergência evolutiva entre sequências completas de cpDNA de Brachiaria e outras gramíneas. As estimativas do erro padrão (s) são mostradas acima da diagonal. Todas as posições que contêm lacunas e dados faltantes no alinhamento foram eliminadas. Um total de 128.636 posições foi considerado no conjunto de dados utilizado na análise. Análises filogenéticas foram realizadas usando o programa MEGA5.

Species	*B.decumbens*	*B.brizantha*	*B.ruziziensis*	*B.humidicola*	*P.virgatum*	*Sorghum bicolor*	*Zea mays*	*Oryza sativa*
B.decumbens		4,16	12,70	38,52	23,67	29,33	33,63	64,22
B.brizantha	130,00		17,83	40,67	23,35	29,49	33,18	68,02
B.ruziziensis	700,00	705,00		22,80	19,98	25,70	37,06	48,97
B.humidicola	1232,00	1220,00	1255,00		23,49	35,46	41,57	54,93
P.virgatum	2072,00	2061,00	2107,00	2199,00		28,57	29,41	38,06
Sorghum bicolor	2959,00	2956,00	2999,00	3080,00	2692,00		22,06	33,36
Zea mays	3141,00	3141,00	3177,00	3255,00	2893,00	1092,00		32,20
Oryza sativa	6590,00	6568,00	6624,00	6701,00	6411,00	6525,00	6701,00	

Conclusões

- As regiões *barcoding* do cpDNA *trn*H-*psb*A, *rbc*L e *mat*K não apresentaram diferenças significativas que dessem suporte a uma análise filogenética e possibilitassem a

diferenciação de quatro espécies de *Brachiaria* (*B. ruziziensis, B. brizantha* e *B. decumbens e B. humidicola)*. Isto se deve à baixa variabilidade de sequência de DNA detectada nestas regiões nas quatro espécies analisadas.

- Na região nuclear 5.8S do rDNA (ITS) foi possível detectar variação de sequência entre espécies de braquiária testadas. Os dados possibilitaram a construção de uma árvore filogenética onde pode ser observado que *B. ruziziensis, B. brizantha* e *B. decumbens* pertencem ao mesmo grupo monofilético. Observou-se ainda que *B. humidicola* é facilmente separado deste grupo. A diferenciação entre *B. ruziziensis, B. brizantha* e *B. decumbens* é muito limitada nesta região, referente a um pequeno número de substituições nucleotídicas (entre 2 e 5 bases), ao contrário de *B. humidicola*, onde o número de substituições detectadas foi bem maior (>50 pb).
- Os resultados de montagem do cpDNA de quatro espécies de *Brachiaria* (*B. ruziziensis, B. brizantha*, *B. decumbens* e *B. humidicola)* usando um genoma de referência (*P. virgatum*) mostraram, inicialmente, que o sequenciamento NGS de DNA total é suficiente para a recuperação de quantidade suficiente de segmentos de leitura exclusivos do cpDNA para a montagem do genoma cloroplástico de cada uma das quatro espécies. Os quatro genomas de cloroplasto montados possuem uma estrutura circular típica, com grande região de cópia única (*Large Single Copy* - LSC) e uma pequena região de cópia única (*Small Single Copy* - SSC), separadas por duas cópias de inversão repetida (*Inverted Repeat* - IR). O tamanho dos genomas do cloroplasto obtidos variaram entre 138.765 bp em *B. ruziziensis* e 138.976 bp em *B. humidicola*.
- Nenhum dos quatro genomas sequenciados cobriu a toda a sequência do genoma referência, o qual apresentou cerca de 800 pb a mais que a sequência do genoma de *Brachiaria ruziziensis,* evidenciando que os genoma de cloroplasto destas espécies de *Brachiaria* parecem ser menores do que a sequência de *Panicum virgatum,* que apresenta 139.619 bases.
- O tamanho final das sequências de cpDNA montados para *B. ruziziensis, B. brizantha, B. decumbens* e *B. humidicola* é muito próximo (138.765, 138.945, 138.940, 138.976). A diferença foi de apenas 5 pb entre *B. brizantha* e *B. decumbens*, e de até 208 bp entre *B. humidicola* e *B. ruziziensis*.
- A montagem de cpDNA com base em pequenos segmentos de leitura de DNA é uma

estratégia desafiadora, mas que se apresenta rápida e eficiente. Os quatro genomas cpDNA foram montados *de novo*. A cobertura linear alcançada neste trabalho pelo somatório das sequências montadas *de novo* (*scaffolds*) variou entre 92,89 a 99,45%. O N50 do tamanho dos contigs resultantes do sequenciamento *de novo* do cpDNA de *B. ruziziensis* (1.704) foi cerca de três vezes maior do que o obtido para os três demais genomas, que tiveram um N50 variando de 485 a 505 bases. Contudo, os maiores contigs montados não foram de *B. ruziziensis,* mas sim os de *B. brizantha* e *B. humidicola*.

- O alinhamento das sequências montadas de cpDNA das quatro espécies possibilitou a seleção de regiões indel que permitem a separação de acessos de cada espécie. Foram selecionados para validação um total de 18 indels que apresentam polimorfismo de inserção/deleção *in silico* e permitem distinguir as quatro espécies de *Brachiaria* (*B. ruziziensis, B. brizantha, B. decumbens e B. humidicola*). Análise laboratorial confirmou a utilidade dos marcadores indels na separação de acessos de diferentes espécies de *Brachiaria*.
- O genoma do cloroplasto das quatro espécies de *Brachiaria* contém 118 genes únicos, dos quais 18 são duplicados nas regiões invertidas IRs, perfazendo um total de 136 genes de função conhecida. Além disso, existem nove ORFs e três pseudogenes.
- Apesar de terem sido detectadas sequências cpSSR no DNA cloroplástico das quatro espécies, o polimorfismo verificado *in silico* nestas regiões é muito limitado.
- Verificou-se um número de SNPs em comparações par-a-par variando de 50 SNPs, existentes entre as sequências de cloroplasto de *B. brizantha* e *B. decumbens*, a 1018 SNPs, identificados na comparação entre as sequencias de *B. decumbens* e *B. humicola*. A menor quantidade de SNPs computados nas comparações par-a-par foi encontrado entre as espécies filogeneticamente mais próximas.
- As indels de um único nucleotídeo (1 base) foram as mais comum nas comparações entre sequências de cpDNA, representando pelo menos 40% do total. As espécies mais próximas apresentam um menor número de indels entre si (ex. *B. brizantha* e *B. decumbens*). No entanto, o número de indels identificadas não diminuiu com o aumento do tamanho da inserção/deleção. Há uma elevada correlação (0,856512) entre o número de indels e o número de SNPs encontrados nas comparações interespecíficas.
- As relações filogenéticas das quatro espécies de *Brachiaria* foram ainda exploradas por

uma análise comparativa das quatro sequências completas do cpDNA (usando referência), juntamente com sequências completas de *Panicum virgatum, Oryza sativa, Zea mays e Sorghum bicolor* do Genbank. O relógio molecular foi calibrado usando o ponto de divergência de *Oryza sativa* e *Zea mays*, com ocorrência estimada há 65 milhões de anos (MYA). Todas as árvores filogenéticas construídas tiveram a mesma topologia e indicam que *B. decumbens* e *B. brizantha* podem ser derivadas de um ancestral comum com *B. ruziziensis.* Indicam ainda que *B. humidicola* é a espécie mais distante nas comparações de polimorfismo de cpDNA com *B. ruziziensis, B. decumbens e B. brizantha*..

- Estima-se que o tempo de divergência entre *B. decumbens* e *B. brizantha* seja apenas 2,5 MYA, e que estas duas espécies tenham se separado do ancestral que originou *B. ruziziensis* há 14 MYA. Isto provoca a hipótese de que *B. decumbens* e *B. brizantha* tenham surgido do ancestral de *B. ruziziensis* por evento(s) de poliploidização, que isolou reprodutivamente estas espécies.

Referências

1. Renvoize, S., et al., *Morfología, taxonomía y distribución natural de Brachiaria (Trin.) Griseb.* Brachiaria: Biología, Agronomía y Mejoramiento. CIAT. Cali, Colombia, 1998: p. 1-17.
2. Maass, B., et al., *Identificación y nomenclatura de las especies de Brachiaria.* Brachiaria: biología, agronomía y mejoramiento, 1998.
3. Loch, D., *Brachiaria decumbens (signal grass): a review with particular reference to Australia.* Trop. Grasslands, 1977. **11**(2): p. 141-157.
4. Palmer, J.D., *Chloroplast DNA Evolution and Biosystematic Uses of Chloroplast DNA Variation.* American Naturalist, 1987. **130**(s1).
5. Clegg, M. and G. Zurawski, *Chloroplast DNA and the Study of Plant Phylogeny: Present Status and Future Prospects*, in *Molecular Systematics of Plants*, P. Soltis, D. Soltis, and J. Doyle, Editors. 1992, Springer US. p. 1-13.
6. Baldwin, B.G., *Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the compositae.* Mol Phylogenet Evol, 1992. **1**(1): p. 3-16.
7. Hamby, R.K. and E. Zimmer, *Ribosomal RNA as a Phylogenetic Tool in Plant Systematics*, in *Molecular Systematics of Plants*, P. Soltis, D. Soltis, and J. Doyle, Editors. 1992, Springer US. p. 50-91.
8. Zimmer, E., et al., *Rapid duplication and loss of genes coding for the alpha chains of hemoglobin.* Proceedings of the National Academy of Sciences, 1980. **77**(4): p. 2158-2162.
9. Birky, C.W., *Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution.* Proceedings of the National Academy of Sciences, 1995. **92**(25): p. 11331-11338.
10. Shinozaki, K., et al., *The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression.* The EMBO journal, 1986. **5**(9): p. 2043.
11. Ohyama, K., et al., *Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA.* 1986.
12. Hiratsuka, J., et al., *The complete sequence of the rice (Oryza sativa) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals.* Molecular and General Genetics MGG, 1989. **217**(2-3): p. 185-194.
13. Wolfe, K.H., C.W. Morden, and J.D. Palmer, *Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant.* Proceedings of the National Academy of Sciences, 1992. **89**(22): p. 10648-10652.
14. Palmer, J.D., et al., *3 Chloroplast and Mitochondrial DNAs of Arabidopsis thaliana: Conventional Genomes in an Unconventional Plant.* Cold Spring Harbor Monograph Archive, 1994. **27**: p. 37-62.
15. Dong, W., et al., *Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding.* PLoS ONE, 2012. **7**(4): p. e35071.
16. Hollingsworth, P.M., S.W. Graham, and D.P. Little, *Choosing and Using a Plant DNA Barcode.* PLoS ONE, 2011. **6**(5): p. e19254.
17. Riaz, T., et al., *ecoPrimers: inference of new DNA barcode markers from whole genome*

sequence analysis. Nucleic Acids Research, 2011. **39**(21): p. e145.
18. Baldwin, B.G., et al., *The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny.* Annals of the Missouri Botanical Garden, 1995: p. 247-277.
19. Shaw, J. and R.L. Small, *Chloroplast DNA phylogeny and phylogeography of the North American plums (Prunus subgenus Prunus section Prunocerasus, Rosaceae).* American Journal of Botany, 2005. **92**(12): p. 2011-2030.
20. Taberlet, P., et al., *Universal primers for amplification of three non-coding regions of chloroplast DNA.* Plant Mol Biol, 1991. **17**(5): p. 1105-9.
21. Group, C.P.W., et al., *A DNA barcode for land plants.* Proceedings of the National Academy of Sciences, 2009. **106**(31): p. 12794-12797.
22. Doyle, J. and J. Doyle, *A rapid DNA isolation procedure for small quantities of fresh leaf tissue*. 1987.
23. White, T., et al., *Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics*, in *PCR Protocols: A Guide to Methods and Applications*, M. Innis, et al., Editors. 1990, Academic Press. p. 315-322.
24. Sang, T., D. Crawford, and T. Stuessy, *Chloroplast DNA phylogeny, reticulate evolution, and biogeography of Paeonia (Paeoniaceae).* American Journal of Botany, 1997. **84**(8): p. 1120-1120.
25. Tate, J.A. and B.B. Simpson, *Paraphyly of Tarasa (Malvaceae) and Diverse Origins of the Polyploid Species.* Systematic Botany, 2003. **28**(4): p. 723-737.
26. Kress, W.J. and D.L. Erickson, *A Two-Locus Global DNA Barcode for Land Plants: The Coding <italic>rbcL</italic> Gene Complements the Non-Coding <italic>trnH-psbA</italic> Spacer Region.* PLoS ONE, 2007. **2**(6): p. e508.
27. Dunning, L.T. and V. Savolainen, *Broad-scale amplification of matK for DNA barcoding plants, a technical note.* Botanical Journal of the Linnean Society, 2010. **164**(1): p. 1-9.
28. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
29. Hall, T.A., *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.* Nucleic Acids Symposium Series, 1999. **41**: p. 95-98.
30. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
31. Felsenstein, J., *Confidence limits on phylogenies: an approach using the bootstrap.* Evolution, 1985: p. 783-791.
32. Ferreira, M.E. and D. Grattapaglia, *Introdução ao uso de marcadores RAPD e RFLP em análise genética*. 1995: Embrapa-Cenargen.
33. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res, 2010. **20**(2): p. 265-72.
34. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
35. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic acids research, 1994. **22**(22): p. 4673-4680.
36. Tamura, K. and M. Nei, *Estimation of the number of nucleotide substitutions in the*

control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol, 1993. **10**(3): p. 512-26.
37. Nei, M. and S. Kumar, *Molecular evolution and phylogenetics*. 2000: Oxford University Press.
38. Muller, K.F., T. Borsch, and K.W. Hilu, *Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting matK, trnT-F, and rbcL in basal angiosperms.* Mol Phylogenet Evol, 2006. **41**(1): p. 99-117.
39. Bortiri, E., et al., *The complete chloroplast genome sequence of Brachypodium distachyon: sequence comparison and phylogenetic analysis of eight grass plastomes.* BMC Research Notes, 2008. **1**(1): p. 61.
40. Kolodner, R. and K. Tewari, *Inverted repeats in chloroplast DNA from higher plants.* Proceedings of the National Academy of Sciences, 1979. **76**(1): p. 41-45.
41. Wang, W. and J. Messing, *High-Throughput Sequencing of Three <italic>Lemnoideae</italic> (Duckweeds) Chloroplast Genomes from Total DNA.* PLoS ONE, 2011. **6**(9): p. e24670.
42. Zhang, W., et al., *A Practical Comparison of <italic>De Novo</italic> Genome Assembly Software Tools for Next-Generation Sequencing Technologies.* PLoS ONE, 2011. **6**(3): p. e17915.
43. Graham, S.W., et al., *Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference.* International Journal of Plant Sciences, 2000. **161**(S6): p. S83-S96.
44. Kelchner, S.A., *The Evolution of Non-Coding Chloroplast DNA and Its Application in Plant Systematics.* Annals of the Missouri Botanical Garden, 2000. **87**.
45. Leseberg, C.H. and M.R. Duvall, *The complete chloroplast genome of Coix lacryma-jobi and a comparative molecular evolutionary analysis of plastomes in cereals.* J Mol Evol, 2009. **69**(4): p. 311-8.
46. Xu, Q., et al., *Analysis of Complete Nucleotide Sequences of 12 <italic>Gossypium</italic> Chloroplast Genomes: Origin and Evolution of Allotetraploids.* PLoS ONE, 2012. **7**(8): p. e37128.
47. Provan, J., et al., *DNA fingerprints of rice (Oryza sativa) obtained from hypervariable chloroplast simple sequence repeats.* Proc Biol Sci, 1996. **263**(1375): p. 1275-81.
48. Wang, Y., et al., *Exploring the Switchgrass Transcriptome Using Second-Generation Sequencing Technology.* PLoS ONE, 2012. **7**(3): p. e34225.
49. Marechal, A. and N. Brisson, *Recombination and the maintenance of plant organelle genome stability.* New Phytol, 2010. **186**(2): p. 299-317.
50. Gray, B.N., B.A. Ahner, and M.R. Hanson, *High-level bacterial cellulase accumulation in chloroplast-transformed tobacco mediated by downstream box fusions.* Biotechnology and bioengineering, 2009. **102**(4): p. 1045-1054.
51. Yamane, K., K. Yano, and T. Kawahara, *Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice.* DNA research, 2006. **13**(5): p. 197-204.
52. Young, H.A., et al., *Chloroplast genome variation in upland and lowland switchgrass.* PLoS One, 2011. **6**(8): p. e23980.
53. Muse, S.V., *Examining rates and patterns of nucleotide substitution in plants.* Plant Mol Biol, 2000. **42**(1): p. 25-43.

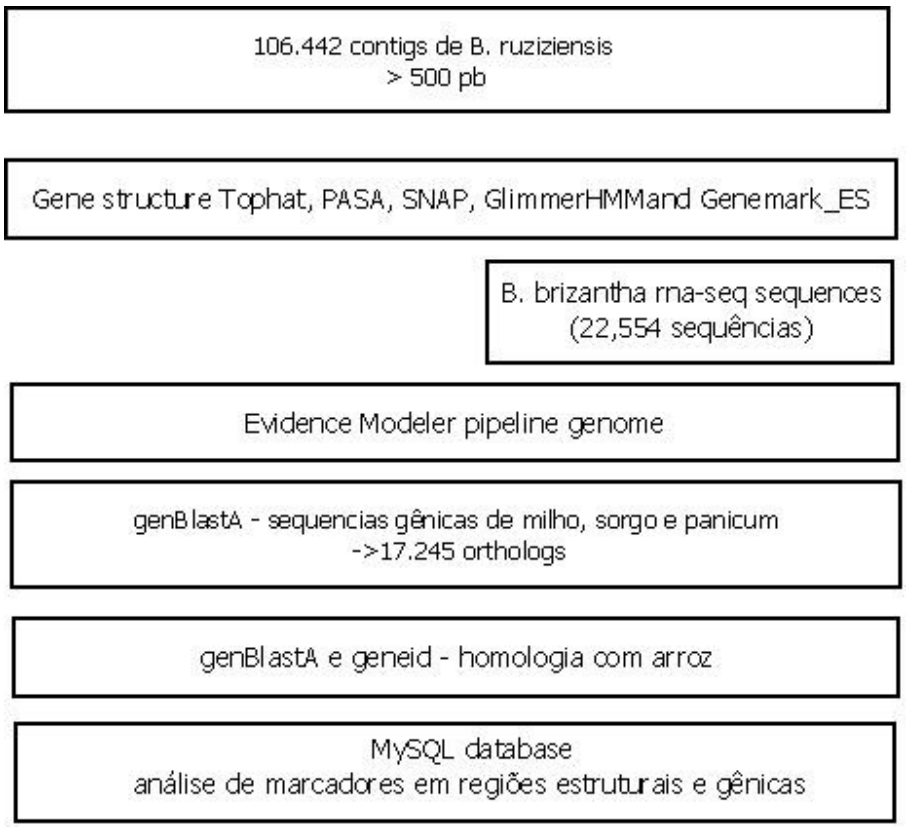
54. Jakobsson, M., et al., *The evolutionary history of the common chloroplast genome of Arabidopsis thaliana and A. suecica.* Journal of evolutionary biology, 2007. **20**(1): p. 104-121.
55. Zhang, W., *Phylogeny of the grass family (Poaceae) from rpl16 intron sequence data.* Mol Phylogenet Evol, 2000. **15**(1): p. 135-46.
56. Group, G.P.W., et al., *Phylogeny and subfamilial classification of the grasses (Poaceae).* Annals of the Missouri Botanical Garden, 2001: p. 373-457.

XI. ANEXOS

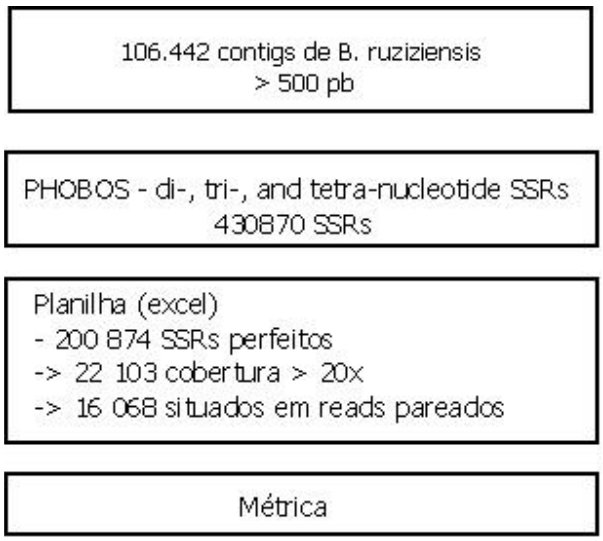
ANEXO 1. Pipeline de montagem *de novo* do genoma nuclear de *B. ruziziensis*; Montagem de novo;

(a) Montagem *de novo*;

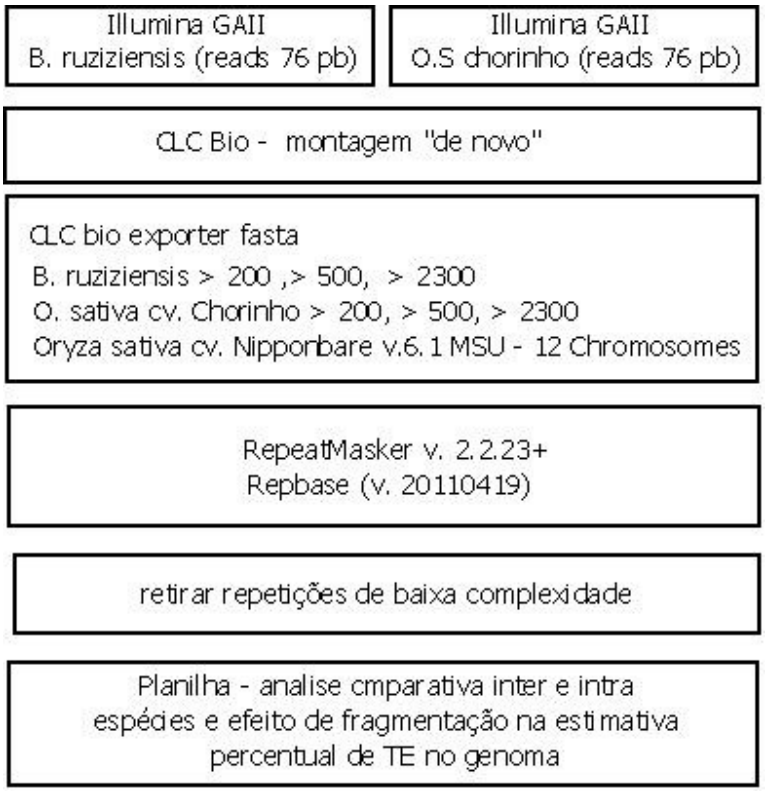
(b) Anotação gênica;



(c) Detecção e desenvolvimento de microssatélites;

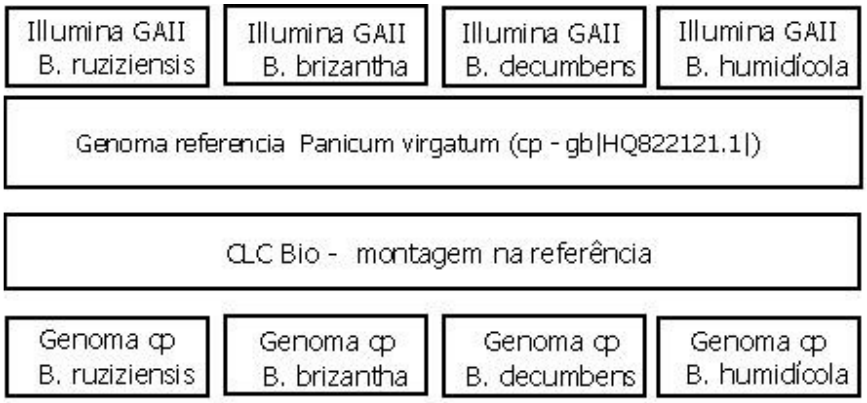


(d) Detecção de Elementos Transponíveis;

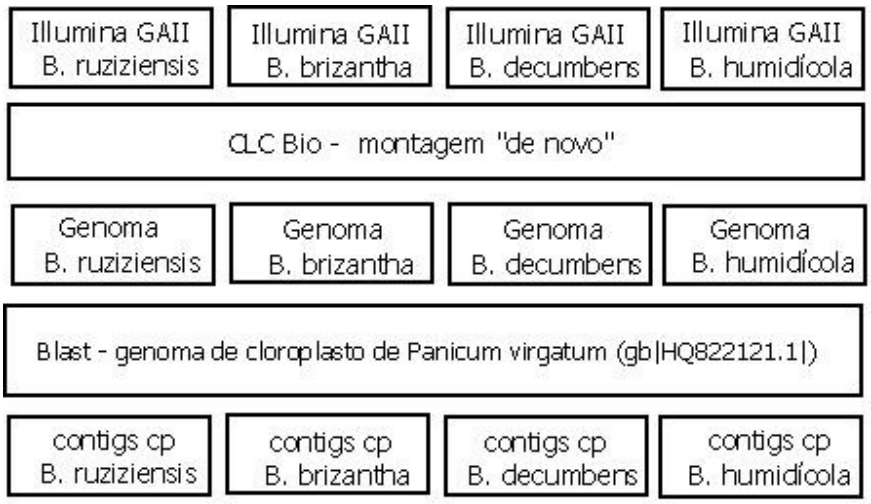


ANEXO 2. Pipeline e montagem do genoma cloroplástico;

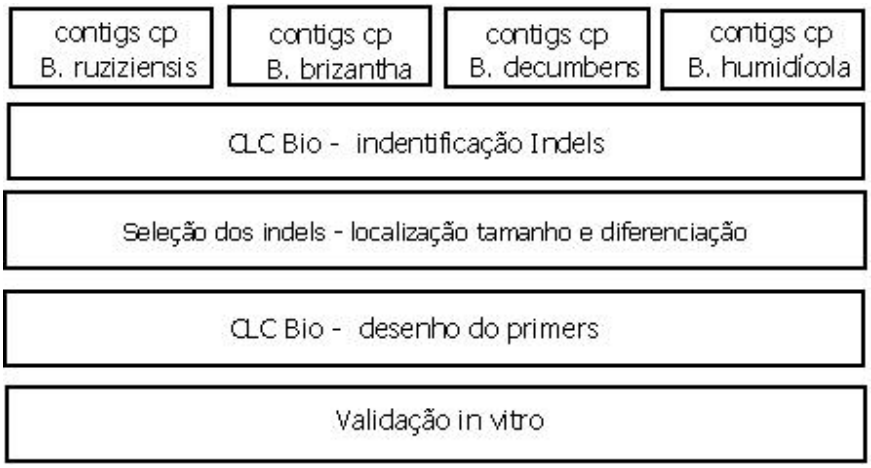
(a) Montagem com genoma referência de *Panicum*;



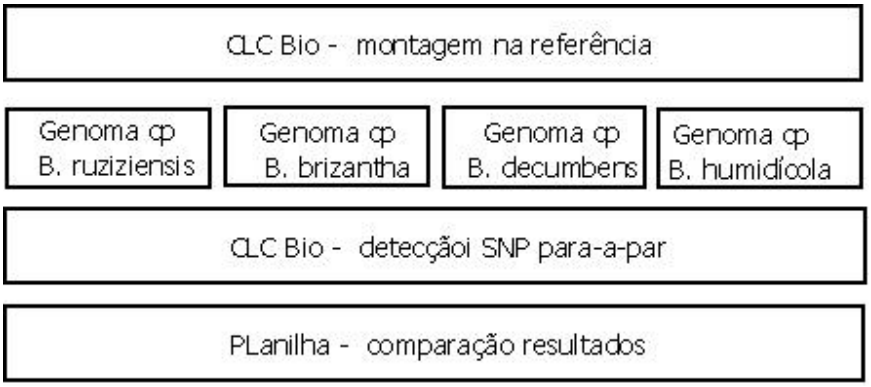
(b) Montagem *de novo*;



(c) Detecção e desenvolvimento de indels;



(d) Detecção e desenvolvimento de SNPs;



ANEXO 3. Tabela com lista de primers de indels

Identifcador	Orientation	tm	Seq	Expected Product size
66584	FORWARD	49.71	AAGAAGTTCTTACTCTTTCTGT	**105**
	REVERSE	49.90	ACATACGACTCATAATGAA	
72645	FORWARD	50.18	GAAAGAGAAAAAAGTTGTC	**153**
	REVERSE	50.73	AGAGTGGATCAAGAAAAAA	
72956	FORWARD	49.77	TCATCTGTCTTTCTTTCC	**175**
	REVERSE	49.13	CTATCAGAAAACCACTAT	
74248	FORWARD	54.85	CGATGCAAAGAAAATGAATG	**119**
	REVERSE	51.71	CGTAAGATCCCATAGAGT	
75494	FORWARD	52.22	AGTTCTCGCTTTAAATCC	**193**
	REVERSE	48.82	CCCTAGATACCTAAAATC	
79281	FORWARD	56.67	GCCCGCGAAATCCTTATT	**162**
	REVERSE	52.90	CAAAACTGGACATGAGAG	
81154	FORWARD	48.37	TGAAGTCAGTAGGAGT	**153**
	REVERSE	48.66	GGAATCGAAATCTTGG	
81616	FORWARD	48.68	AAAGATTCAGAATAAACAAA	**149**
	REVERSE	48.43	GAAGAAGAACGGGCTAAGGAAA	
107669	FORWARD	54.76	CGAGCATCCAAAACCAAAA	**224**
	REVERSE	55.26	ATGGATAACGGAGGGATT	
113003	FORWARD	49.76	CAAGGAAGGAAAAAGATA	**177**
	REVERSE	48.54	AGTAAACTAGACGAAGAA	
114885	FORWARD	49.74	TTTCTAATCCCTCACTAAC	**177**
	REVERSE	49.07	GTAAACATAAGCAGTGTAA	
119374	FORWARD	49.74	CTTCTTCTCCTCAGCCATT	**109**
	REVERSE	49.07	CATCACATCCCCTCTCTC	
103778/11748	FORWARD	53.83	ATTGGATTTGGATAGAAGGGTA	**95**
	REVERSE	54.36	GCAATAAAAAAATCAGCAAAATTC	
86220/135017	FORWARD	52.31	GTTAGATAGGAACAGCTTTG	**121**
	REVERSE	52.90	TTTATGAACGGGAATGGG	
87460/133763	FORWARD	49.69	TAAGTAGCGATCAAGGAA	**123**
	REVERSE	49.90	GCTCAAAGAACGAATAAA	
93252/127974	FORWARD	52.68	TAAGTAGCGATCAAGGAA	**157**
	REVERSE	52.84	GCTCAAAGAACGAATAAA	

ANEXO 4. Relação dos genes identificados no cpDNA de *Brachiaria ruziziensis*.

Gene	start	end	Gene	start	end
atpA	35238	36758	rps12	91342	91581
atpB	53190	54683	rps12_3end	91351	91581
atpE	52780	53190	rps12_3end	127988	128218
atpF	33768	33926	rps14	37358	37666
atpF	34738	35142	rps15	103105	103374
atpH	33050	33292	rps15	116190	116459
atpI	31498	32238	rps16	4430	4648
ccsA	107420	107959	rps18	66097	66585
ccsA	107938	108372	rps19	80837	81115
cemA	59075	59764	rps19	138449	138727
clpP	68129	68776	rps2	30531	31238
infA	76657	76974	rps3	79570	80241
lhbA	11567	11752	rps4	45778	46380
matK	1665	3299	rps7	90255	90722
ndhA	112786	113322	rps7	128847	129314
ndhA	114329	114862	rps8	77057	77464
ndhB	87722	88474	rrn16	93583	95074
ndhB	89185	89961	rrn16	124490	125981
ndhB	129608	130384	rrn23	97494	100381
ndhB	131095	131847	rrn23	119183	122070
ndhC	50420	50779	rrn4.5	100477	100571
ndhD	108597	110096	rrn4.5	118993	119087
ndhE	110936	111238	rrn5	100799	100919
ndhF	103488	105701	rrn5	118645	118765
ndhG	111437	111964	trnA-UGC	96467	96504
ndhH	114873	116051	trnA-UGC	97315	97349
ndhI	112150	112689	trnA-UGC	122215	122249
ndhJ	49098	49574	trnA-UGC	123060	123097
ndhK	49680	50426	trnC-GCA	19048	19118
orf188	114317	114859	trnD-GUC	16102	16175
orf42	96648	96764	trnE-UUC	15564	15636
orf42	122800	122916	trnF-GAA	48445	48517
orf56	96965	97057	trnfM-CAU	12380	12453
orf56	97054	97083	trnfM-CAU	37135	37188
orf56	97068	97133	trnG-UCC	12047	12117
orf56	122431	122496	trnH-GUG	81245	81319
orf56	122481	122510	trnH-GUG	138245	138319
orf56	122507	122599	trnI-CAU	83334	83407

petA	60005	60964	trnI-CAU	136157	136230
petB	72473	73126	trnI-GAU	95378	95419
petD	74010	74531	trnI-GAU	96367	96401
petG	64236	64346	trnI-GAU	123163	123197
petL	63962	64054	trnI-GAU	124145	124186
petN	18036	18128	trnK-UUU	1375	1407
psaA	40060	42309	trnK-UUU	3886	3923
psaB	37830	40031	trnL-CAA	87073	87153
psaC	110219	110461	trnL-CAA	132416	132496
psaI	57685	57792	trnL-UAA	47546	47580
psaJ	65142	65273	trnL-UAA	48122	48171
psbA	88	1146	trnL-UAG	107263	107342
psbC	9526	10983	trnM-CAU	14987	15045
psbD	8556	9614	trnM-CAU	52592	52664
psbE	62389	62637	trnN-GUU	101505	101576
psbF	62260	62376	trnN-GUU	117988	118059
psbH	71376	71594	trnP-GGG	64689	64759
psbI	7203	7355	trnP-UGG	64687	64761
psbJ	61873	61992	trnQ-UUG	6265	6337
psbK	6685	6867	trnR-ACG	101178	101251
psbL	62121	62234	trnR-ACG	118313	118386
psbM	17222	17323	trnR-UCU	36900	36971
psbN	71144	71272	trnS-GCU	7479	7566
psbT	70956	71063	trnS-GGA	45414	45500
psi_psbT	69274	70791	trnS-UGA	11138	11225
rbcL	55428	56858	trnT-GGU	14981	15052
rpl14	77605	77973	trnT-GGU	15051	15115
rpl16	78084	78485	trnT-GGU	52598	52656
rpl2	81378	81806	trnT-UGU	46702	46774
rpl2	82467	82856	trnV-GAC	93288	93359
rpl2	136708	137097	trnV-GAC	126205	126276
rpl2	137758	138186	trnV-UAC	51722	51758
rpl20	66815	67171	trnV-UAC	52361	52399
rpl22	80302	80748	trnW-CCA	64477	64550
rpl23	57128	57286	trnY-GUA	15698	15781
rpl23	82881	83159	ycf1	101901	102056
rpl23	136405	136683	ycf1	102049	102135
rpl32	106553	106741	ycf1	102640	102735
rpl33	65620	65817	ycf1	116829	116924
rpl36	76442	76552	ycf1	117429	117515
rpoA	74752	75768	ycf1	117508	117663
rpoB	20142	23366	ycf15	85994	86290
rpoC1	23407	25455	ycf15	133279	133575

rpoC2	25637	30241	ycf2	83611	83988
rps11	75836	76264	ycf2	134033	134152
rps12	67868	67981	ycf2	135576	135953
ycf3	42946	43101	ycf68	95519	95809
ycf3	43836	44063	ycf68	95811	95921
ycf3	44806	44937	ycf68	123643	123753
ycf4	58154	58708	ycf68	123755	124045