Hinc patriam sustinet

Instituto Superior de Agronomia
Universidade de Lisboa

Identificação de marcadores SSR e de SNPs em medronheiro (Arbutus unedo L.) por sequenciação massiva paralela

Pedro Miguel Jesus Fazenda

Dissertação para obtenção do Grau de Mestre em

Engenharia Agronómica

Orientador: Doutor José Manuel Peixoto Teixeira Leitão

Coorientador: Doutora Cristina Maria Moniz Simões Oliveira

Júri:

Presidente: Doutor José Luís Monteiro Teixeira, Professor Associado do Instituto Superior de Agronomia da Universidade de Lisboa.

Vogais: Doutor José Manuel Peixoto Teixeira Leitão, Professor Catedrático da Faculdade de Ciências e Tecnologia da Universidade do Algarve.

Doutora Maria Leonor Mota Morais Cecílio, Professora Auxiliar do Instituto Superior de Agronomia da Universidade de Lisboa.

2013

U LISBOA | UNIVERSIDADE DE LISBOA

Agradecimentos

Gostaria de expressar aqui o meu agradecimento a todos aqueles que tornaram possível a realização deste trabalho.

Ao meu orientador, Professor Doutor José Leitão, por ter aberto as portas do seu laboratório e pelo incentivo, disponibilidade e apoio sempre demonstrados em todas as fases da realização deste trabalho.

Aos meus colegas de laboratório, em especial ao Doutor Jorge Carlier pelo apoio, incentivo e sugestões, que em muito contribuiram para a realização deste trabalho.

Aos meus pais, que sempre me incentivaram e me permitiram chegar a esta etapa.

A todos os meus amigos, em especial àqueles com quem convivi durante a minha vida académica na Universidade do Algarve e no Instituto Superior de Agronomia, e que me apoiaram de diversas formas.

Resumo

O medronheiro *(Arbutus unedo* L.) é uma espécie autóctone da região mediterrânea. O uso de marcadores moleculares nesta espécie tem sido limitado ao uso de RAPDs, ISSRs e ainda à amplificação cruzada com SSRs provenientes de outras *Ericaceae*. Neste trabalho, desenvolveu-se um protocolo de extração de DNA nuclear de medronheiro e procedeu-se à sequenciação massiva paralela parcial do genoma de *Arbutus unedo* L. na plataforma "Ion Torrent" (Life Technologies). Obtiveram-se 198.856 sequências ("raw data") com um tamanho médio de 123 bp, disponibilizadas na base de dados "Sequence Read Archive" (SRA) do NCBI com o número de acesso: SRX341237. A análise destes dados levou à identificação de 1085 sequências com motivos microssatélite, as quais foram disponibilizadas na base de dados de sequências nucleotídicas do NCBI com os números de acessão: de KF023636 a KF024720. Desenhou-se primers para 18 loci microssatélites dos quais unicamente 3 se revelaram polimórficos num painel de 16 amostras. Foram identificados 25 SNPs e foi desenvolvido um marcador CAPS, que apesar de ser heterozigótico se revelou monomórfico nas 16 amostras analisadas.

Palavras-chave: *Arbutus unedo*, CAPS, extração de DNA nuclear, marcadores moleculares, sequenciação massiva paralela, SSRs.

Abstract

The strawberry tree (*Arbutus unedo* L.) is native to the Mediterranean region. The use of molecular markers in this species has been limited to the use of RAPDs, ISSRs as well to the cross-amplification of SSRs from other *Ericaceae*. In this work, we developed a protocol for extracting nuclear DNA from the strawberry tree and performed partial next-generation sequencing of the *Arbutus unedo* L. genome using the "Ion Torrent" (Life Technologies) platform. The next-generation sequencing resulted in 198,856 sequences ("raw data") with an average size of 123 bp, which were uploaded to the NCBI database "Sequence Read Archive" (SRA) with the accession number: SRX341237. Data analysis led to the identification of 1085 microsatellite-containing sequences, which were also uploaded with accession numbers: from KF023636 to KF024720 to the NCBI databases. Primers were designed for 18 microsatellite loci of which only three have proved to be polymorphic in a panel of 16 samples. Based on identified 25 SNPs one CAPS marker was developed, which despite being heterozygous revealed to be monomorphic among the 16 analyzed samples.

Keywords: *Arbutus unedo*, CAPS, molecular markers, next-generation sequencing, nuclear DNA extraction, SSRs.

Extended Abstract

Strawberry tree (*Arbutus unedo* L.) is a member of the Ericaceae family spread all over the Mediterranean region and quite common in the Algarve region, mainly in "Serra de Monchique" and "Serra do Caldeirão". The fruits are used by local inhabitants for the production of cakes, jams and especially spirit drinks. "Medronho", the distilled liquor obtained from fruits is the major source of income from this species, since the fresh fruit consumption is restricted to local populations during the harvest season.

Strawberry tree has not been subject of intensive plant breeding programs aimed to obtain high quality fruit producing cultivars although prospection work for selection of genotypes producing high quality fruits have been performed in Italy and Turkey (Mulas *et al*., 1998; Celikel *et al*., 2008; and Sulusoglu *et al*., 2011). Conversely, other genera of the Ericaceae family as *Vaccinium,* the most important genus for berry production in North America, have been subject of intensive and extensive plant breeding and of multiple genetic diversity studies including the use of molecular markers techniques (Boches *et al*., 2006; Bassil *et al.,* 2010; Zhu *et al.,* 2011)

In *A. unedo,* genetic diversity studies using RAPD (Random Amplified Polymorphic DNA) markers have been performed recently in Portugal (Gomes *et al*., 2012; Lopes *et al*., 2012) and in Tunisia (Takrouni and Boussaid, 2010). Low level of genetic diversity within and between the Portuguese strawberry tree populations was observed by Lopes *et al*., (2012), which was also confirmed by Gomes *et al*., (2012) who found an average genetic similarity of about 0.83 between individuals within populations with some genotypes sharing up to 95% of the amplified bands. It is noteworthy that these results are similar to those obtained in Tunisia by Takrouni and Boussaid (2010), though the slightly higher genetic differentiation between the Tunisian populations.

Microssatellites or SSR (Single Sequence Repeat) markers have been developed for other Ericaceae, in particular for *Vaccinium macrocarpon* Ait. (Zhu *et al.,* 2011) and *Vaccinium corymbosum* L. (Boches *et al*., 2006). Even though Gomes *et al*. (2012) have tested SSR markers developed for *Vaccinium* for cross amplification in strawberry trees at the best of my knowledge, so far, specific microsatellite markers have not been developed for this last species.

The main goal of the present work was the identification by next-generation sequencing (or massively parallel sequencing) of a large number of SSR sequences

in *A. unedo*. The strong DNA degrading activity found in leaf homogenates has required a novel procedure for extraction of large amounts of high quality nuclear DNA to be previously established.

The initial tests for DNA extraction with different protocols, including protocols routinely used in the LGGI (Laboratory of Genomics and Genetic Improvement, FCT, UAlg) for DNA extraction in multiple plant species have resulted in highly degraded DNA. To overcome this problem a novel procedure involving previous isolation of nuclei was developed and immediately adapted for use with small samples in eppendorf tubes. An improved protocol based on that created by Doyle and Doyle (1987) allowed the extraction of large amounts of high quality DNA from isolated nuclei. The excellent amplifiability of the extracted DNA was tested by RAPD analysis.

The preparation of the DNA for sequencing was performed according to the instructions described in the "User Bulletin - *Preparing Short Amplicon (<250 bp) Libraries Using the Ion Plus Fragment Library Kit*" published by Life Technologies. The massively parallel sequencing carried out in an Ion PGM™ sequencer resulted in 198,856 sequences with an average size of 123 bp. These sequences were uploaded as a Sequence Read Archive (SRA) with the accession number SRX341237 to the NCBI databases (http://www.ncbi.nlm.nih.gov/). Among these sequences, 1085 were identified as containing microsatellite repeats and uploaded to the NCBI databases with the accession numbers from KF023636 to KF024720.

A microsatellite sequence was identified per 10,21 Kb. Dinucleotides represented 95% of the repeated sequences. The dinucleotide AG constituted 72% of all microsatellite sequences and the dinucleotide CG was the less frequent. This differential frequency seems to be specific to the Ericaceae family since results have been previously found in *Vaccinium macrocarpon* (Zhu *et al*., 2011). Primers were designed for 18 microsatellite loci, but only 3 out of them were found polymorphic among 16 *A. unedo* plants, which suggest a very low genetic diversity within the analyzed population.

Ten sequences containing 25 SNPs (Single Nucleotide Polymorphisms) were also identified. Primers were designed for one of these sequences which was converted into a CAPS (Cleaved Amplified Polymorphic DNA) marker. The analysis among the same 16 DNA samples revealed that all analyzed plants shared the same, monomorphic, heterozygous restriction pattern.

Although the study of genetic diversity is out of the scope of the present dissertation, the few data obtained indicate a very low level of genetic variability among the analyzed 16 strawberry tree plants. Although confirming the previous observations of other authors with other populations, the assessed low level of diversity is unexpected for a seed propagated putatively allogamous species as *A. unedo*.

Keywords: *Arbutus unedo*, CAPS, molecular markers, next-generation sequencing, nuclear DNA extraction, SSRs

Índice

Lista de Quadros

Lista de Figuras

Lista de Abreviaturas

bp – “base pair”
CAPS – “Cleaved Amplified Polymorphic Sequence”
CTAB – “Cetyl trimethyl ammonium bromide”
DAPI – “4’,6-Diamidino-2-Phenylindole”
DNA – “Deoxyribonucleic Acid”
dNTPs – “Deoxynucleotide Triphosphates”
DTT – “Dithiothreitol”
EDTA – “Ethylenediamine tetraacetic acid”
EST – “Expressed Sequence Tag”
IBA – “Indol Butyric Acid”
kb – “kilobase”
LGMG – Laboratório de Genómica e Melhoramento Genético
PCR – “Polymerase Chain Reaction”
PMSF – “Phenylmethylsulfonil fluoride”
PVP – “Polyvinyl pyrrolidone”
RAPD – “Random Amplified Polymorphic DNA”
SDS – “Sodium docedyl sulfate”
SNP – “Single Nucleotide Polymorphism”
SSR – “Single Sequence Repeat”
TE- “Tris-EDTA”
Tris – “Tris(hydroxymethyl)aminomethane”
UV – “Ultraviolet”

Introdução

O medronheiro

O medronheiro (*Arbutus unedo* L.) pertence à família *Ericaceae*, e ocorre espontâneamente como um arbusto ou pequena árvore que geralmente não ultrapassa os 5 m, podendo chegar aos 10 m. É de folhagem perene e as suas folhas são alternas, oblongo-lanceoladas, margem serrada a sub-inteira e pecíolo curto. As flores, reunidas em panículas terminais pendentes, são urceoladas, de cor branca, esverdeada ou rosada (Bingre *et al*., 2007). Os seus frutos são bagas globosas, de 2-3 cm verdes/amareladas no estado imaturo e vermelhas quando maduras. O amadurecimento dos frutos e o aparecimento das flores, ocorre simultaneamente e dá-se de Outubro até Fevereiro. É uma espécie predominantemente autógama, embora também possa ocorrer polinização anemófila e entomófila (Hagerup, 1957). Este mesmo autor, relata nas suas observações de medronheiros cultivados em estufa aquecida na Dinamarca, o vingamento de frutos na ausência de insetos e ainda a forte adaptação da biologia e morfologia floral para a anemofilia e não para a entomogamia. O método mais vulgar de propagação é por semente, embora não seja fácil devido à necessidade de quebra de dormência (Tilki, 2004; Ertekin e Kirdar, 2010), que por vezes pode ser difícil e demorada. Também é possível a propagação do medronheiro por estacaria, no entanto os resultados publicados variam muito, pois Mereti *et al*. (2002), referem percentagens de enraizamento *ex vitro* inferiores a 60 % e Sulusoglu (2012) obteve 100 % de estacas enraizadas numa das modalidades testadas. Este último autor conclui ainda que tanto o genótipo como a concentração da hormona usada (IBA) têm grande influência na capacidade de enraizamento.

É uma planta interessante do ponto de vista ecológico, pois promove a biodiversidade da fauna, já que o seu fruto é alimento para várias aves e a manta morta é abrigo para insetos que também servem de alimento a aves. Quando forma povoamentos dominantes, a projeção da sua copa serve de abrigo a pequenos mamíferos. O seu sistema radicular muito ramificado conjugado com a manta morta protegem o solo da erosão. É também resistente a condições de temperatura extrema, à seca e ao fogo, rebentando vigorosamente após um incêndio. Distribui-se maioritariamente por toda a bacia do mediterrâneo e em Portugal encontra-se frequentemente associado ao sobreiro e ao pinheiro-bravo, raramente formando povoamentos dominantes (Pedro, 1994). Cresce em solos siliciosos, bem como em descarbonatados, numa gama de pH que pode variar entre 5 a 7.2 (Torres *et al.*, 2002; Celikel *et al.*, 2008). É uma espécie bastante comum no interior Algarvio (Serras de Monchique e do Caldeirão) e sempre se revelou de grande importância para a economia das populações locais, que utilizam o fruto para fabrico de bebidas destiladas (aguardente),

compotas e bolos e também pelo fato das flores do medronheiro serem uma importante fonte de néctar para as abelhas, ainda mais, numa época do ano em que poucas espécies estão em floração. Em Portugal, uma grande parte da produção de fruto é destinada para o fabrico de aguardente. O consumo do fruto em fresco é um hábito apenas entre as populações locais, sobretudo durante a jornada de colheita, embora actualmente, não só em Portugal, haja um interesse crescente em divulgar esta forma de consumo do fruto, pois é uma boa fonte de compostos com propriedades antioxidantes, como flavonóides (Males *et al*., 2006), antocianinas (Pawlowska *et al.*, 2006), bem como de vitamina C e betacaroteno (Alarcão-e-Silva *et al*., 2001).

O medronheiro é uma espécie sub-explorada em Portugal e noutros países mediterrânicos. Os frutos, têm sido recolhidos, para propagação em viveiro, de indíviduos que crescem espontâneamente no campo, dando pouca ou nenhuma atenção à qualidade do material usado. Embora as plantas daqui obtidas sejam à partida heterogéneas, este método funcionou durante muito tempo para satisfazer as necessidades dos viveiristas que faziam revenda para "garden centers". Com o interesse dos profissionais da fileira florestal em clones de medronheiro mais produtivos, começou então a ser dada mais importância à propagação vegetativa desta espécie, sendo exemplos disso os trabalhos em Portugal de Gomes *et al*. (2012) e na Turquia de Sulusoglu (2012).

Relativamente à seleção de clones, o medronheiro raramente foi alvo de intensos programas de melhoramento para obter cultivares com frutos de alta qualidade (Celikel *et al*., 2008). No entanto, além dos estudos enunciados anteriormente acerca dos constituintes do fruto, encontram-se também alguns estudos sobre os constituintes químicos das folhas que indicam que a presença de um elevado teor de fenóis em extratos aquosos e etanólicos (Oliveira *et al.*, 2009) bem como de α-tocoferol (Kivçak e Mert, 2001) que lhes confere uma forte atividade antioxidante, possibilita o uso desta espécie nas indústrias química e farmacêutica. Os poucos trabalhos de seleção/prospeção de campo de genótipos com qualidade de fruto superior foram feitos em Itália, onde Mulas *et al*. (1998) avaliaram as características do fruto de 20 genótipos da Sardenha, e sobretudo na Turquia, onde Celikel *et al*. (2008) e Sulusoglu *et al*. (2011), seleccionaram várias plantas de qualidade superior (com maior peso do fruto, maior teor de sólidos solúveis, entre outras) para posterior propagação vegetativa e disponibilização para cultivo extensivo.

Pelo contrário, outros géneros da família Ericaceae, têm sido objecto de intensos programas de melhoramento, incluindo estudos de diversidade genética associados ao uso e desenvolvimento de marcadores moleculares, como é o caso do género de grande

importância ornamental, *Rhododendron* (Wang *et al.*, 2010) e do género mais importante na produção de pequenos frutos na América do Norte, *Vaccinium* (Boches *et al*., 2006; Bassil *et al.*, 2010; Zhu *et al.*, 2011). Estes aspetos, apenas começaram a ser abordados em *A.unedo* em estudos recentes. Dão conta disso, os estudos de diversidade genética usando a técnica RAPD (Random Amplified Polymorphic DNA) feitos em Portugal (Gomes *et al*., 2012; Lopes *et al*., 2012) e na Tunísia (Takrouni e Boussaid, 2010). Lopes *et al*. (2012) analisaram 46 genótipos do interior Norte e Centro de Portugal (distritos de Viseu, Vila Real, Bragança e Castelo Branco) e encontraram níveis baixos a moderados de diversidade genética dentro das populações, sendo a população de Bragança, a população com área de distribuição mais ampla a que registou maior diversidade. Entre populações o nível de diferenciação também foi baixo. Os mesmos autores, comparam ainda estes resultados com os obtidos por Takrouni e Boussaid (2010) onde a diversidade genética dentro das populações foi semelhante, mas a diferenciação entre populações foi ligeiramente superior. Gomes *et al*. (2012) analisaram 9 populações (Gerês, duas em Coimbra, Serra do Açor, Serra da Gardunha, Serra de Alvélos, Serra da Arrábida e no Algarve – São Marcos da Serra e Serra do Caldeirão) e apesar de terem encontrado polimorfismos com os primers RAPD usados, a análise de agrupamento calculada com o coeficiente de Lynch (Lynch, 1990) revelou uma similaridade de 83 % entre indivíduos, sendo que alguns genótipos partilharam até 95 % das bandas.

Além dos marcadores moleculares RAPD, os marcadores microssatélites ou SSR (Single Sequence Repeat) também já foram desenvolvidos e utilizados no género *Ericaceae,* sobretudo em *Rhododendron delavayi* Franch. (Wang *et al.*, 2010), em *Vaccinium macrocarpon* Ait., conhecido como arando ou “cranberry” (Zhu *et al.*, 2011) e *Vaccinium corymbosum* L., conhecido como mirtilo (Boches *et al*., 2006). Até à data, e pelo nosso conhecimento, não há registo de publicações que relatem o desenvolvimento de microssatélites em *A. unedo*, no entanto devido à alta reprodutibilidade destes marcadores entre laboratórios e pelo fato de serem tranferíveis entre espécies ou mesmo géneros estreitamentes relacionados Gomes *et al*. (2012) utilizaram em *A. unedo* onze pares de primers SSR desenvolvidos em *Vaccinium* por Boches *et al*. (2005), dos quais, nove produziram produtos amplificados em todos os genótipos testados e cinco foram polimórficos. Á semelhança dos resultados anteriormente obtidos com RAPDs (Gomes *et al*., 2012) os microssatélites também não permitiram a Gomes *et al*. (2012) agrupar os genótipos analisados de acordo com a sua origem geográfica. Apesar disto, reforçou-se a ideia da amplificação cruzada na avaliação da diversidade genética entre géneros diferentes (*Arbutus* e *Vaccinium*) e foi dado um passo importante no estabelecimento de um conjunto

de marcadores aptos a distinguir clones de medronheiro, aspeto este, importante no melhoramento da espécie.

Marcadores SSR e CAPS (Cleaved Amplified Polymorphic Sequence)

Os métodos clássicos de caracterização e identificação das plantas são baseados no fenótipo, uma abordagem que pode ser influenciada pelas condições ambientais e pelo longo tempo de geração.

Polimorfismos genéticos como os encontrados em sequências microssatélite e os SNPs (Single Nucleotide Polymorphisms) são abundantes nos genomas de plantas e constituem importantes ferramentas na identificação de genótipos. Os microssatélites são pequenas sequências de DNA (1-6 bp) repetidas em tandem que se encontram espalhados por todo o genoma e são geralmente hipervariáveis, exibindo um grande número de alelos em cada locus (Litt e Luty 1989; Tautz, 1989; Weber e May 1989). São por isso largamente utilizados em estudos genéticos como por exemplo, em testes de paternidade. Os SNPs, consistem na alteração de um nucleótido específico numa sequência de DNA e representam a forma mais frequente de variação do DNA no genoma (Brookes, 1999). Uma das formas de detetar SNPs é sob a forma de marcador CAPS (Cleaved Amplified Polymorphic Sequence) por corte com uma enzima de restrição específica cujo sítio de corte foi criado ou eliminado pelo SNP. À semelhança dos microssatélites, os marcadores CAPS podem ser analisados em gel de agarose ou poliacrilamida após uma reação de PCR (Polymerase Chain Reaction) seguida do corte com a enzima de restrição desejada (Konieczny e Ausubel, 1993).

Métodos clássicos de identificação de SSRs e SNPs

Um dos inconvenientes dos microssatélites é a necessidade da sua identificação prévia que pode ser morosa e difícil. Os métodos tradicionais de identificação de microssatélites baseiam-se no rastreio de bibliotecas genómicas com sondas apropriadas (Rassmann *et al*. 1991) cujo sucesso varia entre 12 % a <1 % de clones positivos (geralmente entre 2-3 % para muitos grupos taxonómicos), na hibridização seletiva de fragmentos de DNA com sondas contendo repetições e posterior amplificação de PCR do DNA enriquecido (Karagyozov *et al*., 1993; Armour *et al*., 1994; Kijas *et al*.,1994) e na hibridização por "Southern blotting" de perfis RAPD com sondas contendo repetições seguida da clonagem seletiva de bandas positivas (Ender *et al*., 1996). Informações detalhadas sobre estes e outros métodos de identificação de microssatélites podem ser encontradas na revisão feita por Zane *et al*. (2002).

A identificação de SNPs tem sido feita através da sequenciação de genes ou fragmentos importantes de um certo número de indivíduos ou linhas, e posterior comparação das sequências obtidas para identificar os SNPs. Este método é moroso, caro e não se adapta à necessidade de obter um número elevado de SNPs. Devido ao número crescente de projetos de sequenciação de genomas de plantas, ficaram disponíveis ao público, em bases de dados, um grande número de sequências genómicas de sequências expressas ou EST (Expressed Sequence Tag), o que tem permitido a descoberta *in silico* (simulação em computador) de SNPs em muitas novas espécies através da pesquisa de bases de dados e posterior validação por PCR (Batley *et al*., 2003).

Sequenciação Massiva Paralela

Desde que apareceram no mercado, em 2005, as novas tecnologias de sequenciação massiva paralela, conhecida também como "next-generation sequencing" tiveram um grande impacto nos estudos de genómica. A sua aceitação foi imediata, devido às enormes vantagens que trouxe comparativamente ao método de Sanger. As várias plataformas disponíveis no mercado, como a '*454 FLX*' da Roche, a '*Solexa*' da Illumina, a '*SOLiD System*' da Applied Biosystems e a '*Ion Torrent*' da Life Technologies, continuam atualmente a sofrer melhoramentos e muitos softwares têm sido desenvolvidos para lidar com o crescente número de dados. Estas plataformas são capazes de gerar informação sobre milhões de sequências numa única corrida e com uma grande economia de tempo e do custo por base que teriam com o método Sanger (Shendure e Ji, 2008). Para contornar os problemas do método Sanger algumas plataformas de próxima geração, como a '454 FLX' e a 'Ion Torrent' usam um método de amplificação de DNA *in vitro* conhecido como, PCR em emulsão (Tawfik *et al*., 1998). Neste método, fragmentos individuais de DNA são ligados a adaptadores marcados com biotina e depois capturados individualmente em esferas em emulsão. Cada gotícula de emulsão (contendo uma esfera) age como um reactor de amplificação individual, produzindo milhões de cópias de um único padrão de DNA.

O método usado pela "Ion Torrent" baseia-se na deteção da alteração de pH pela libertação de iões de hidrogénio que acompanha a incorporação de novos nucleótidos durante a síntese de DNA. Ao contrário de outras plataformas, o sistema da "Ion Torrent" não utiliza detetores óticos, e ao não usar fluorescência ou quimiluminescência, reduz drasticamente o custo de sequenciação.

Uma das aplicações atuais e emergentes da sequenciação massiva paralela é a identificação em larga escala de polimorfismos genéticos (Shendure e Ji, 2008), em

particular de SSRs, como são exemplos os trabalhos de Zhu *et al*. (2011) e Georgi *et al*. (2012) em *Vaccinium macrocarpon* e SNPs, como é exemplo o trabalho de Yang *et al*. (2013) em *Lupinus angustifolius*. Tanto os marcadores SSR como CAPS usam primers específicos nas amplificações de PCR, o que implica um conhecimento prévio, ainda que reduzido, do genoma, conhecimento esse que não existe em medronheiro, pois até à data não existem dados publicados.

O objectivo deste trabalho, foi dar um contributo nesta área inexplorada, fazendo a primeira sequenciação parcial do genoma de *A. unedo* com recurso às novas tecnologias de sequenciação massiva paralela e identificar marcadores SSR. Caraterísticas enzimáticas intrínsecas do medronheiro obrigaram ao desenvolvimento prévio de um protocolo eficiente para a extração em larga quantidade de DNA genómico nuclear de boa qualidade.

Materiais e Métodos

Material Vegetal

Para a extração de DNA e posterior sequenciação foi colhido de cada planta - um pequeno ramo lenhoso com folhas saudáveis. Colheu-se amostras de uma planta no *campus* de Gambelas (Gb), Faro e de quinze plantas da empresa Cortevelada, Odiáxere (Od), Lagos. A extração de DNA realizou-se no próprio dia ou no dia seguinte à recolha do material.

Extração de DNA

Inicialmente foram testados dois protocolos de extração de DNA anteriormente usados neste laboratório, Laboratório de Genómica e Melhoramento Genético (LGMG). O tampão de extração de DNA consistia para o protocolo 1 em: 300 mM Tris HCl pH 8.0, 25 mM EDTA pH 8.0, 250 mM NaCl, 1 % SDS e 2 % PVP; e para o protocolo 2 em: 300 mM Tris HCl pH 8.0, 25 mM EDTA pH 8.0, 2 M NaCl, 2 % CTAB e 2 % PVP. Foram efetuadas variações a estes protocolos, nomeadamente na concentração de NaCl, na concentração e momento de aplicação de RNase-A e purificação com diferentes solventes orgânicos.

Quantificação do DNA

A integridade e concentração aproximada do DNA foi verificada, por comparação com DNA de *Pisum* de concentração conhecida, submetendo a amostra a electroforese em gel de agarose (1.2 %) a uma diferença de potencial de 5 V/cm. Os géis foram analisados, sob transiluminação com radiação UV, após imersão em solução de brometo de etídeo e fotografados com uma câmara digital “Kodak EDAS 120”. O DNA foi também quantificado num espectrofotómetro “NanoDrop 2000” (Thermo Scientific), sendo depois diluído em água Milli-Q para uma concentração apropriada (2.5 ng/µL) para usar em reações de amplificação por PCR.

Amplificações RAPD

A análise RAPD foi usada para verificar a qualidade em termos de amplificação do DNA extraído pelos diferentes métodos testados. Um total de 13 primers da Operon Technologies Inc. (designados OP) foi usado para testes de amplificação: AA01, AA02, AA03, AK01, AK02, AK03, AK04, AK05, AK06, AK07, AK08, AK09 e AK10. As reações de PCR foram feitas num volume final de 30 µL, contendo 10 ng de DNA genómico, 1x de DreamTaq™

Green Buffer, 0.16 mM de dNTPs, 1.2 U de NZYTaq DNA Polimerase, 0.66 µM de primer e água Milli-Q até perfazer o volume. A amplificação de PCR foi efectuada num Termociclador Biometra UNO II (Thermoblock, Biotron) como a seguir descrito: desnaturação inicial a 94 ºC durante 1 min. e 30 seg, seguido por 35 ciclos de 30 seg. a 94 ºC, 30 seg. a 36 ºC, 1 min. a 72 ºC e um ciclo final de 10 min. a 72ºC. Os produtos de amplificação foram separados por electroforese em gel de agarose (2 %) a uma diferença de potencial de 5 V/cm. Os géis foram analisados, sob transiluminação com radiação UV, após imersão em solução de brometo de etídeo e fotografados com uma câmara digital "Kodak EDAS 120".

Análise de marcadores microssatélite (SSR)

Dezoito pares de primers foram desenhados com base em sequências microssatélite identificadas por sequenciação massiva paralela. A síntese dos primers foi realizada pela *NZYTech Lda.-Genes & Enzymes.* As reações de PCR foram efetuadas num volume final de 15 µL de composição idêntida à preconizada para as reações RAPD. A amplificação de PCR foi efectuada num termociclador UnoCycler (VWR) de forma idêntica à efetuada para as amplificações RAPD, com excepção da temperatura de "annealing", a qual variou entre 57 e 60 ºC dependendo do par de primers. Os produtos de amplificação foram separados por electroforese em gel de agarose (3 %) a uma diferença de potencial de 5 V/cm e em gel de poli-acrilamida (10 %) a uma diferença de potencial de 13 V/cm, em ambos os casos analisados, sob transiluminação com radiação UV, após imersão em solução de brometo de etídeo e fotografados como acima descrito.

Análise de marcadores CAPS

A análise CAPS foi efetuada usando um par de primers, sintetizados pela *NZYTech Lda.-Genes & Enzymes*. A enzima de restrição usada foi a DraI (Fermentas, Life Sciences). As reações de PCR foram realizadas num volume final de 25 µL de composição idêntica à preconizada para as reações RAPD. A amplificação de PCR foi efectuada num termociclador UnoCycler (VWR) de forma idêntica à efetuada para as amplificações RAPD, com excepção da temperatura de "annealing", que foi de 58 ºC. Os produtos de amplificação de PCR foram analisados por electroforese em gel de agarose (3 %). O corte com a enzima de restrição foi realizado num volume de reação de 15 µL, a 1 x de tampão da enzima, 2 U da enzima, 8 µL do produto amplificado e água Milli-Q até perfazer o volume. A reação de restrição incubou durante 2h30m a 37 ºC, seguida de inactivação a 65 ºC durante 20 min. Os produtos dos cortes foram analisados por electroforese em gel de agarose (2 %), como acima descrito.

Sequenciação Massiva Paralela

A preparação da biblioteca de DNA para sequenciação fez-se submetendo a amostra de DNA do indivíduo Gb a reação de restrição com duas enzimas, Csp6I e TasI (Fermentas, Life Sciences). O volume final de reação foi de 60 µL, contendo 1 x de tampão da enzima, 2 µg de DNA, 20 U de cada enzima e água Milli-Q até perfazer o volume. A reação de restrição incubou durante 10 h, 5h a 37 ºC (Csp6I), seguidas de 5h a 65 ºC (TasI). A separação dos fragmentos foi feita por electroforese em gel de agarose (1.4 %) tendo-se confirmado que a maior parte dos fragmentos tinha cerca de 200 bp. O DNA foi diretamente purificado a partir da mistura de restrição com o kit "GeneJET Gel Extraction Kit" (Fermentas, Life Sciences). Para evitar a contaminação da amostra com tiocianato de guanidina (presente no "binding buffer") alterou-se ligeiramente o protocolo de extração, fazendo-se duas lavagens em vez de uma com 700 µL de tampão e deixando-se atuar durante 5 min. antes das centrifugações. No final, os fragmentos de DNA foram eluídos em 60 µL de "Low" TE do kit "Ion Plus Fragment Library kit" (Life Technologies) e quantificados num espectrofotómetro "NanoDrop 2000" (Thermo Scientific).

A correção dos términos, ligação dos adaptadores e o "nick repair" fez-se com os reagentes do "Ion Plus Fragment Library Kit" (Life Technologies) e o adaptador "Barcode 4" do kit "Ion Xpress Barcode Adapters 1-16". Seguiram-se os procedimentos do "User Bulletin" -> "*Preparing Short Amplicon (<250 bp) Libraries Using the Ion Plus Fragment Library Kit*" (Life Technologies) a partir do passo "End repair" com 100 ng de fragmentos de DNA, tendo a biblioteca sido amplificada por 8 ciclos de PCR. Os fragmentos, que com os adaptadores ligados ficaram com cerca de 300 bp, foram purificados em gel de agarose o que permitiu também eliminar os dímeros de adaptadores (82, 84 e 86 bp). As purificações dos fragmentos a partir da fatia de gel foram sempre realizadas com o kit "GeneJET Gel Extraction Kit" com as modificações descritas anteriormente.
Para se estimar a concentração (ng/µL) da biblioteca de fragmentos de DNA utilizou-se o fluorómetro Qubit ® 2.0 (Life Technologies) com o kit "Qubit ® dsDNA HS Assay Kit" (Life Technologies). A conversão da concentração em ng/µL para ρM, fez-se com base no fator de conversão para fragmentos de DNA de dupla cadeia de 300 bp (5.48 ρmol/1 µg de DNA), como sugerido pela Life Technologies (http://www.invitrogen.com/site/us/en/home/References/Ambion-Tech-Support/rna-tools-and-calculators/dna-and-rna-molecular-weights-and-conversions.html). Em seguida diluiu-se a biblioteca para uma concentração de 26 ρM e com 20 µL desta procedeu-se à ligação dos fragmentos de DNA às nanoesferas, à sua amplificação, nas esferas, por PCR em emulsão e enriquecimento da bilioteca em esferas positivas, *i.e.*, ligadas a fragmentos de DNA amplificados, recorrendo-se ao sistema

automatizado “Ion One Touch System” com o kit “Ion One Touch 200 Template Kit” (Life Technologies). Finalmente a biblioteca enriquecida com esferas positivas foi carregada no sequenciador “Ion PGM™” (Life Technologies) num “chip” Ion 314™ e sequenciada utilizando o kit “Ion PGM™ 200 Sequencing Kit” (Life Technologies).

Análise de dados

As sequências resultantes da sequenciação (“raw data”) foram automaticamente gravadas num ficheiro com formato “.fastq”. A partir deste ficheiro foi possível fazer o controlo de qualidade das leituras, bem como o alinhamento *de novo* das mesmas. Para testar a qualidade individual de cada base utilizou-se a ferramenta “FastQC v. 0.10.0” (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Esta ferramenta produz gráficos de fácil interpretação para analisar as sequências no que diz respeito à qualidade e frequência de tamanhos. Depois de analisada a qualidade da leitura das sequências ao longo da sua extensão e a distribuição de frequências por tamanhos estas foram selecionadas com base no tamanho e aparadas no término com base na qualidade média das bases por posição recorrendo-se para tal, à plataforma “online” “Galaxy” (Giardine *et al.*, 2005; Blankenberg *et al*., 2010a; Goecks *et al*., 2010). As sequências foram filtradas por tamanho, usando a ferramenta “Filter FastQ v. 1.0.0” (Blankenberg *et al.,* 2010b), tendo sido selecionadas as sequências maiores que 100 bp, tamanho mínimo desejado para os produtos de amplificação dos microssatélites. Para aparar as sequências utilizou-se a ferramenta “Trim sequences v. 1.0.0” (http://hannonlab.cshl.edu/fastx_toolkit/). Ainda na plataforma “Galaxy” utilizou-se a ferramenta “FASTQ to FASTA v.1.0.0” (Blankenberg *et al*., 2010b) para criar uma cópia das sequências no formato “.fasta”, *i.e*., sem a informação da qualidade de cada base. Em seguida, utilizou-se a ferramenta “cd-hit-est” da plataforma “online”, CD-HIT Suite (Li *et al*., 2001; 2002; 2006; Ying *et al*., 2010) para eliminar sequências redundantes. Foi utilizado um nível de similaridade limite de 90 % (definido por defeito), ou seja, o programa alinhou sequências com 90 % ou mais de similaridade entre si e escolheu a mais representativa. Jennings *et al*. (2011) utilizaram uma estratégia semelhante para remover sequências redundantes e posterior identificação de SSRs. Deste alinhamento resultou um ficheiro em formato “.fasta”, utilizado para a identificação de sequências microssatélites.

Identificação de sequências SSR e desenho de primers

A identificação de microssatélites foi efetuada no programa “MsatCommander v. 0.8.2” (Faircloth, 2008) colocando como “input” o ficheiro anterior (sem sequências redundantes), e dando ordem para deteção de todos os microssatélites com 6 ou mais repetições de di-, tri-

tetra-, penta- e hexanucleótidos. As sequências contendo microssatélites foram depois alinhadas com as sequências "raw data" utilizando a aplicação "GS Reference Mapper" (disponível no programa Newbler v.2.6, http://454.com/products/analysis-tools/gs-de-novo-assembler.asp). Desta forma foi possível escolher os alinhamentos com maior profundidade (maior número de leituras por base) para desenhar os primers, diminuindo a probabilidade de erros de sequenciação. Os parâmetros usados para o alinhamento foram os que estavam por defeito, excepto o "Mininum overlap length" e o "Minimum overlap identity" que foram 25 e 80 % respectivamente, ou seja, o programa só alinhou sequências que tivessem uma sobreposição ≥ 25 bases e uma similaridade, nessas bases, ≥ a 80 %.

Os primers das sequências microssatélites foram desenhados manualmente segundo os critérios de qualidade: tamanho aproximado entre 18 e 22 bases, % GC > 40, temperatura de "melting" (Tm) ~ 50 ºC, as últimas 3 bases serem diferentes entre si, sendo a última G ou C e não formação de dímeros. Quando não foi possível compatibilizar todos os critérios, deu-se prioridade à não formação de dímeros e aceitou-se temperaturas de "melting" um pouco superiores, entre 51.5-52.5ºC. A avaliação destes critérios fez-se recorrendo ao programa "FastPCR v. 4.0.27" (Kalendar, 2007). Para verificar se os primers amplificavam produtos em sequências distintas daquelas para as quais foram desenhados recorreu-se à ferramenta "primer search v. 5.0.0" (Blankenberg *et al.*, 2007) disponível na plataforma Galaxy, utilizando-se como "inputs" um ficheiro com os primer's desenhados e o ficheiro que se obteve na plataforma CD-HIT Suite.

Identificação de SNPs e desenho de primers

Para a identificação de SNPs foi feito um alinhamento *de novo* no programa "GS *De Novo* Assembler" (disponível no software Newbler v.2.6, http://454.com/products/analysis-tools/gs-de-novo-assembler.asp). O alinhamento partiu do "raw data" e os parâmetros utilizados foram os estabelecidos por defeito, excepto as opções "Large or complex genome" e "Heterozigotic mode", as quais se seleccionaram. Os "contigs" (sequências consenso formadas por sobreposição de sequências) resultantes deste alinhamento, foram visualizados no software Tablet v.1.13.05.17 (Milne *et al.*, 2010) (disponível em http://bioinf.hutton.ac.uk/tablet) e através da exploração manual de vários "contigs", foram identificados os SNPs. Os "contigs" explorados foram os de maior comprimento, pois pretendia-se para a análise CAPS, produtos com pelo menos 400 bp. A seleção de SNPs obedeceu a que as variantes dos nucleótidos alterados se apresentassem na proporção aproximada de 50/50, de modo a evitar os erros de sequenciação. Como ao longo dos grandes contigs existem várias zonas de fratura, zonas em que a continuidade da sequência

do contig se faz apenas pela sobreposição de uma ou duas sequências, foram cuidadosamente escolhidos os locais a partir dos quais e até onde se desenhariam os primers. Os SNPs candidatos a CAPS, foram analisados juntamente com as respetivas regiões flanqueantes no software "NEBcutter v.2.0" (disponível em http://tools.neb.com/NEBcutter2/) para identificar pontos de corte por enzimas de restrição. Para o desenho dos primers procedeu-se tal como para os microssatélites.

Resultados e Discussão

Capítulo 1 – Otimização do protocolo de extração de DNA

Testes preliminares

Nos testes iniciais aos dois protocolos, o tampão de extração adicionado ao macerado das folhas resultou sempre em DNA degradado (Figura 1), sendo o pior resultado obtido com o protocolo 1 (tampão de extração com SDS). Para solucionar este problema procurou-se evitar a atividade das DNases extraíndo o DNA diretamente de núcleos, o que adicionalmente assegurou que na sequenciação não teriam representatividade os DNA cloroplastídico e mitocondrial.

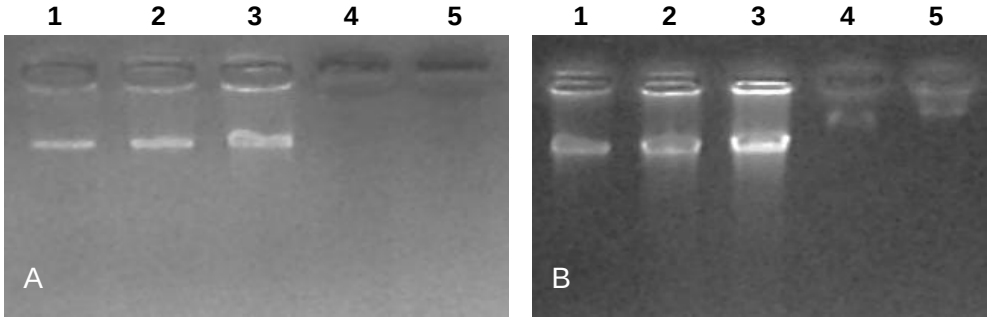

Figura 1. Gel de agarose (1 %). A: Poço 1, 2 e 3 – 100, 250 e 500 ng de DNA de *Pisum*; Poço 4 e 5 – DNA extraído com o protocolo 1. B: Poço 1, 2 e 3 – 250, 500, 1000 ng de DNA de Pisum; Poço 4 e 5 – DNA extraído com o protocolo 2.

Extração de núcleos

Tanto quanto me é permitido saber, o isolamento de núcleos de medronheiro nunca foi feito, como tal, seguiu-se um protocolo usado anteriormente no Laboratório de Genómica e Melhoramento Genético (LGMG) para extração de núcleos de *Pisum*, testando algumas variações no tampão de extração (Quadro 1).

Quadro 1. Tampões de isolamento de núcleos testados em *A. unedo*.

Tampão inicial	50 mM Tris HCl pH 8.0; 0.1 mM PMSF; 15 mM $MgCl_2$; 100 mM NaCl; 0.1 mM $CaCl_2$; 1 mM Spermine; 5 mM DTT; 1 M Sacarose; 0.8 % Triton X-100
Tampão modificado 1	50 mM Tris HCl pH 8.0; 0.1 mM PMSF; 15 mM $MgCl_2$; 100 mM NaCl; 0.1 mM $CaCl_2$; 1 mM Spermine; 5 mM DTT; 0.25 / 0.5 M Sacarose; 1.2 / 1.6 % Triton X-100

Tampão modificado 2	50 mM Tris HCl pH 8.0; 1 M Sacarose; 1.6 % Triton X-100; 15 mM $MgCl_2$
Tampão modificado 3	50 mM Tris HCl pH 8.0; 1 M Sacarose; 1.6 % Triton X-100; 7.5 / 30 mM $MgCl_2$
Tampão modificado 4	50 mM Tris HCl pH 8.0; 1 M Sacarose; 1.6 % Triton X-100; 250 / 500 / 750 mM KCl
Tampão modificado 5	50 mM Tris HCl pH 8.0; 1 M Sacarose; 2 % Triton X-100

O tampão para extração de núcleos 5 (50 mM Tris HCl pH 8.0, 1 M Sacarose e 2 % Triton X-100) em que foram retirados vários componentes presentes no tampão inicial e foi aumentada a concentração de Triton X-100, foi o que apresentou melhores resultados (Fig 2).

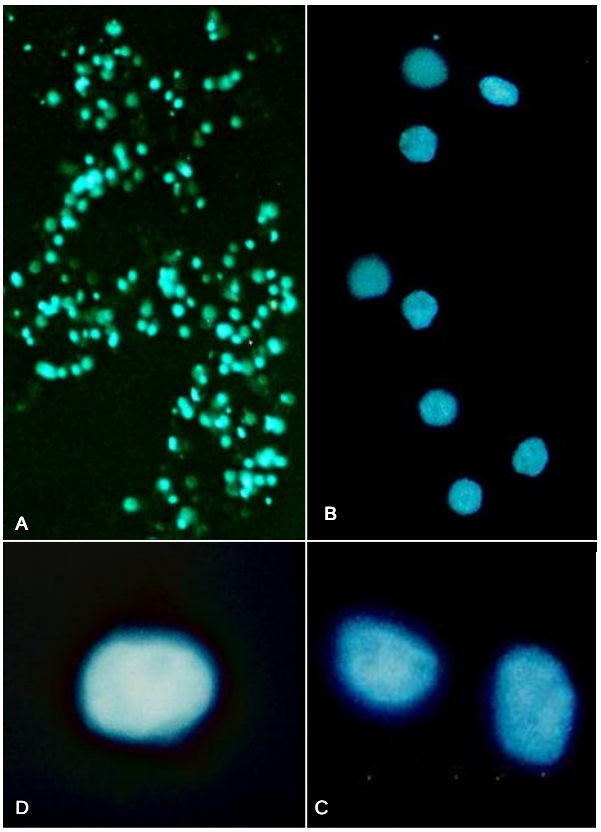

Figura 2. Núcleos de medronheiro corados com DAPI e fotografados num microscópico ótico sob luz UV, com uma câmera fotográfica analógica Olympus. Ampliações: A-500x; B-1000x; C-2000x; D-5000x.

O procedimento a seguir descrito foi otimizado para microtubos de 2 mL, de modo a poder trabalhar com maior número de amostras. O isolamento de núcleos foi avaliado tanto em folhas jovens como em folhas adultas.

Protocolo de extração de núcleos

Macerou-se as folhas (aprox. 75 mg) num almofariz, na presença de azoto líquido, até obter um pó muito fino. Colocou-se o macerado (duas a três pontas de espátula) num tubo de 2 mL, contendo 1 mL de tampão de extração de núcleos (previamente arrefecido em gelo). Centrifugou-se a 80 G durante 2 minutos a 4º C. Com o auxílio de micropipeta com pontas cortadas de forma a aumentar o diâmetro da abertura, retirou-se o sobrenadante para novo tubo tendo o cuidado de não transferir os detritos de maior dimensão. Centrifugou-se a 900 G durante 5 minutos a 4 ºC e descartou-se o sobrenadante, reservando o "pellet" no gelo.

O tampão de extração de núcleos 5 foi aplicado com sucesso a diferentes indivíduos, épocas do ano e estados de expansão das folhas. A maceração das folhas e o rendimento e qualidade dos núcleos extraídos foi sempre ligeiramente melhor para as folhas jovens, visto que nas folhas adultas com maior quantidade de compostos fenólicos e metabolitos secundários (Males *et al.,* 2006; Fiorentino *et al.*, 2007) houve tendência para agregação dos núcleos a restos de cloroplastos, parede celular e outras impurezas.

Extração de DNA

Nos testes iniciais à extração de DNA foi possível observar que o protocolo mais adequado seria o protocolo 2, baseado em Doyle e Doyle (1987). Este protocolo foi o posteriormente utilizado com maior sucesso, alterando a concentração de NaCl de 2 M para 1 M.

O tampão optimizado para a extração de DNA consistiu em 300 mM Tris HCl pH 8.0 (ACROS Organics), 25mM EDTA pH 8.0 (Sigma), 1 M NaCl (Sigma), 2 % CTAB (Sigma) e 2 % PVP (Sigma). Antes de usar adicionou-se proteinase-K (Sigma) a 265 µg/mL e RNase-A (Sigma) a 20 µg/mL.

Ao "pellet" enriquecido em núcleos, juntou-se imediatamente 750 µL de tampão de extração de DNA a quente (previamente aquecido a 75 ºC) e homogeneizou-se rapidamente utilizando um Vortex. Colocou-se a incubar durante 10 min. num banho-maria a 75 ºC agitando ocasionalmente. Juntou-se 1 volume de clorofórmio:álcool isoamílico (24:1), homogeneizou-se muito bem e centrifugou-se a 13.000 rpm durante 5 min. O sobrenadante foi transferido para novo tubo e repetiu-se o passo anterior duas vezes. Ao último sobrenadante recolhido juntou-se 1 volume de isopropanol a -20 ºC, para precipitar o DNA, e homogeneizou-se bem. Deixou-se pelo menos 45min/1hora a -20 ºC. Centrifugou-se a 13.000 rpm durante 5 minutos para recolher o DNA precipitado. O isopropanol foi decantado

e lavou-se o pellet com 300 µL de etanol a 70 % a -20 ºC. Centrifugou-se a 13.000 rpm durante 5 min e retirou-se com uma micropipeta o máximo possível de etanol. Para recolher no fundo do tubo o etanol que ainda ficou nas paredes fez-se um "quick run" de 15 seg. Mais uma vez, retirou-se com uma micropipeta o etanol e deixou-se o tubo aberto para evaporar na sua totalidade. Finalmente, o "pellet" foi ressuspendido em 30 µL de TE (10/0.1) e guardado a 4 ºC

O DNA obtido (Figura 3) com o protocolo proposto, a uma concentração média de 150 ng/µL, foi suficiente para sequenciação massiva paralela e inúmeras amplificações de PCR. Este protocolo diferencia-se do publicado por Sá *et al*. (2011) por evitar o uso de fenol, DTT e 2-mercaptoetanol, compostos potencialmente perigosos para o utilizador. O uso de um composto antioxidante adsorvente, o PVP, provou ser essencial para eliminar os polifenóis, que caso contrário degradariam o DNA (Peterson *et al.*, 1997). Para remover polissacáridos, que futuramente poderiam inibir a atividade da Taq polimerase e de enzimas de restrição (Fang *et al*., 1992), uma quantidade moderada de NaCl, 1 M, mostrou-se suficiente. Purificações com fenol e/ou fenol/clorofórmio/álcool isoamílico (25:24:1) não produziram resultados satisfatórios, pois o rendimento em DNA foi inferior ao obtido com as purificações com clorofórmio/álcool isoamílico (24:1). Para evitar contaminação das amostras com RNA, a adição de RNase-A mostrou ser mais eficaz quando adicionada ao tampão de lise dos núcleos. Um passo importante para garantir a integridade do DNA foi o ajuste da temperatura de lise dos núcleos para 75 ºC. Incubações a temperaturas inferiores resultaram em DNA degradado e contaminado, provavelmente devido às fortes DNases endógenas.

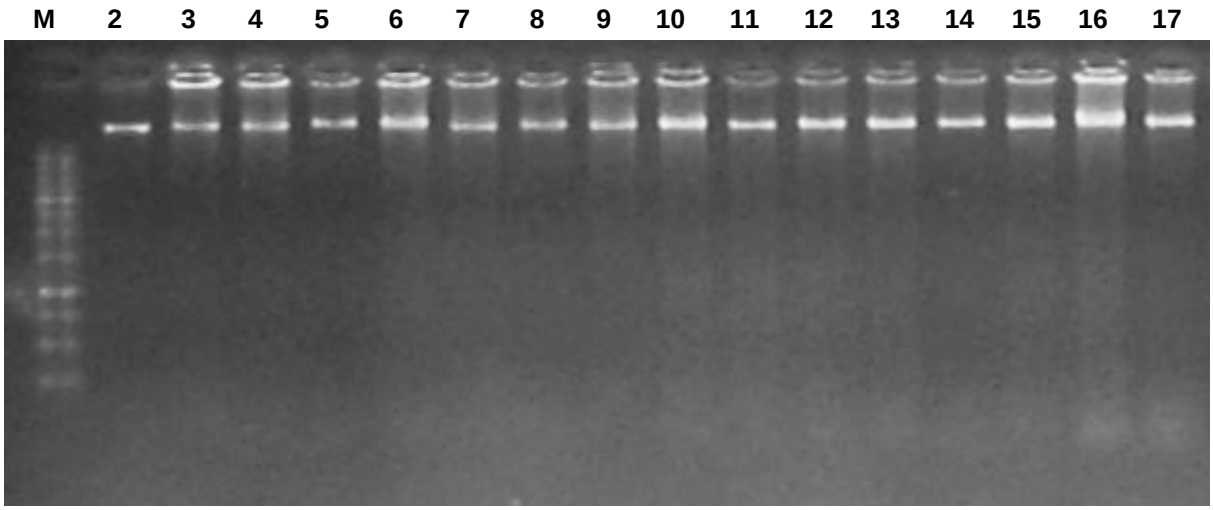

Figura 3. Gel de agarose (1.2 %); M – Marcador de peso molecular "1 kb DNA Ladder" (Fermentas), Poço 2 – 500 ng de DNA de *Pisum*, Poço 3 a 17 – DNA extraído de 15 amostras de *A. unedo* recolhidas em Odiáxere.

Análise qualitativa do DNA por RAPDs

Após quantificação e diluição do DNA, a qualidade amplificativa do DNA foi testada com 13 primers RAPD. Estas análises serviram também para uma primeira apreciação da variabilidade genética das amostras recolhidas .

Os padrões claros e reprodutíveis (não apresentados) dos marcadores RAPD confirmaram a boa qualidade do DNA para amplificação por PCR. O primer OP AA02 resultou no padrão mais polimórfico (Figura 4) e uma breve análise à variabilidade molecular entre indivíduos, sugeriu pouca diversidade na população amostrada.

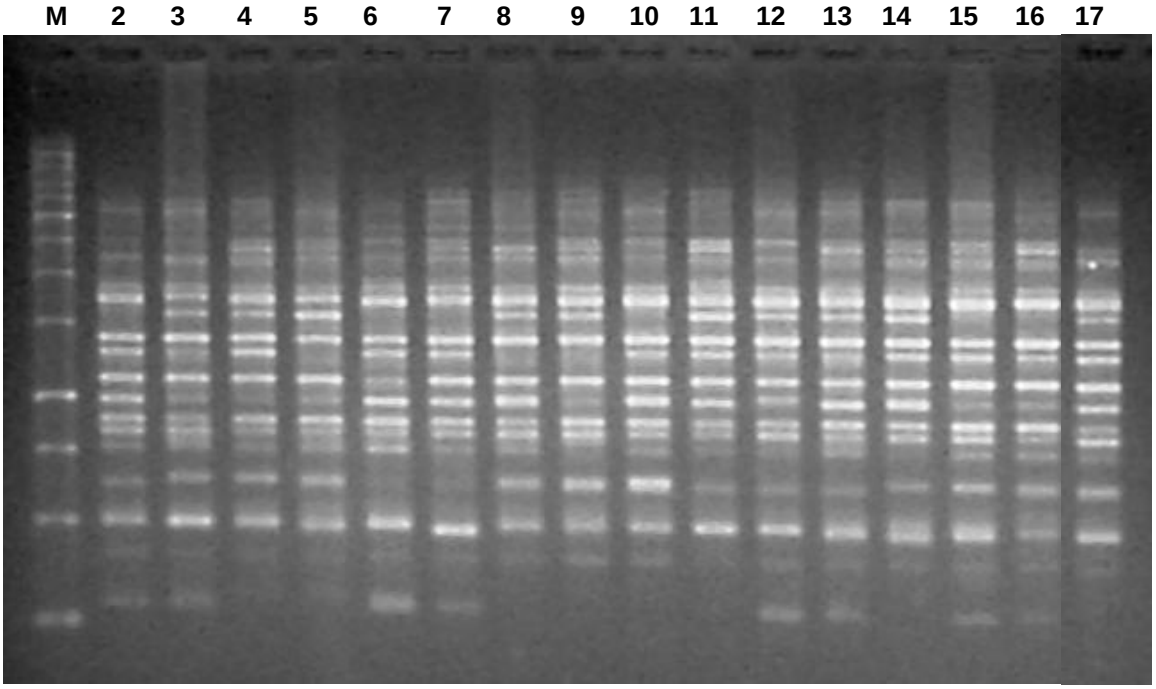

Figura 4. Gel de agarose (2 %). Perfis RAPD do primer OP AA02. M – Marcador de peso molecular “1 kb DNA Ladder” (Fermentas), Poço 2 a 16 – 15 amostras de *A. unedo* de Odiáxere, Poço 17 – amostra de *A. unedo* de Gambelas.

Capítulo 2 – Sequenciação massiva paralela e identificação de sequências SSR e de SNPs

Da sequenciação resultaram 19.5 Mb de informação (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=search_obj&m=&s=&term=SRX341237&go=Search), correspondentes a 198.856 sequências (“raw data”) contendo 24.554.094 de bases e com um comprimento médio de 123 bp (Figura 5).

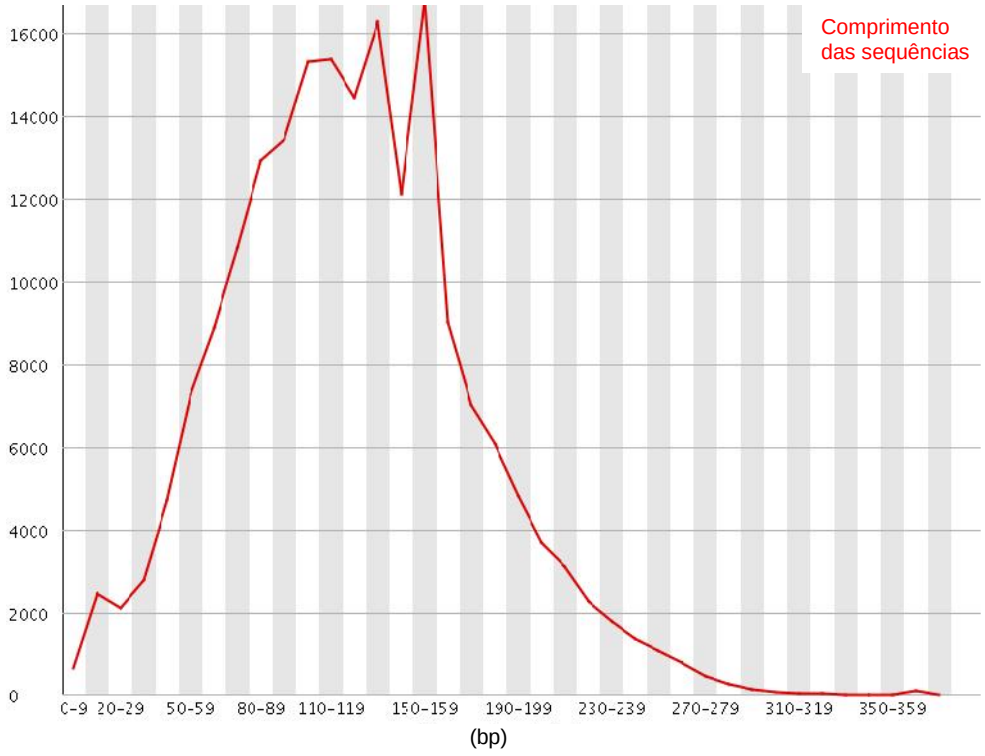

Figura 5. Distribuição de sequências por comprimento.

Identificação de sequências SSR

Para a posterior identificação de motivos microssatélites os “raw data” foram inicialmente trabalhados na plataforma Galaxy, onde após filtragem das leituras maiores de 100 bp restaram 132.681 sequências (Quadro 2). Fez-se de seguida a análise da qualidade média das bases lidas, optando-se por aparar as sequências pela 125ª base, tendo em conta a diminuição do valor da qualidade (Q) a partir aproximadamente dessa posição (Figura 6). Em seguida a redundância destas sequências foi eliminada na plataforma CD-HIT Suite tendo o número de sequências baixado para 99.786.

Quadro 2. Tratamento dos dados para identificação de SSRs.

1ª seleção	Número total de bases (bp)	16.087.869
	Número total de sequências > 100 bp	132.681
	Comprimento médio das sequências (bp)	121
2ª seleção	Número total de bases (bp)	12.104.078
	Número de sequências representativas para uma similaridade de 90%	99.786
	Comprimento médio das sequências (bp)	121

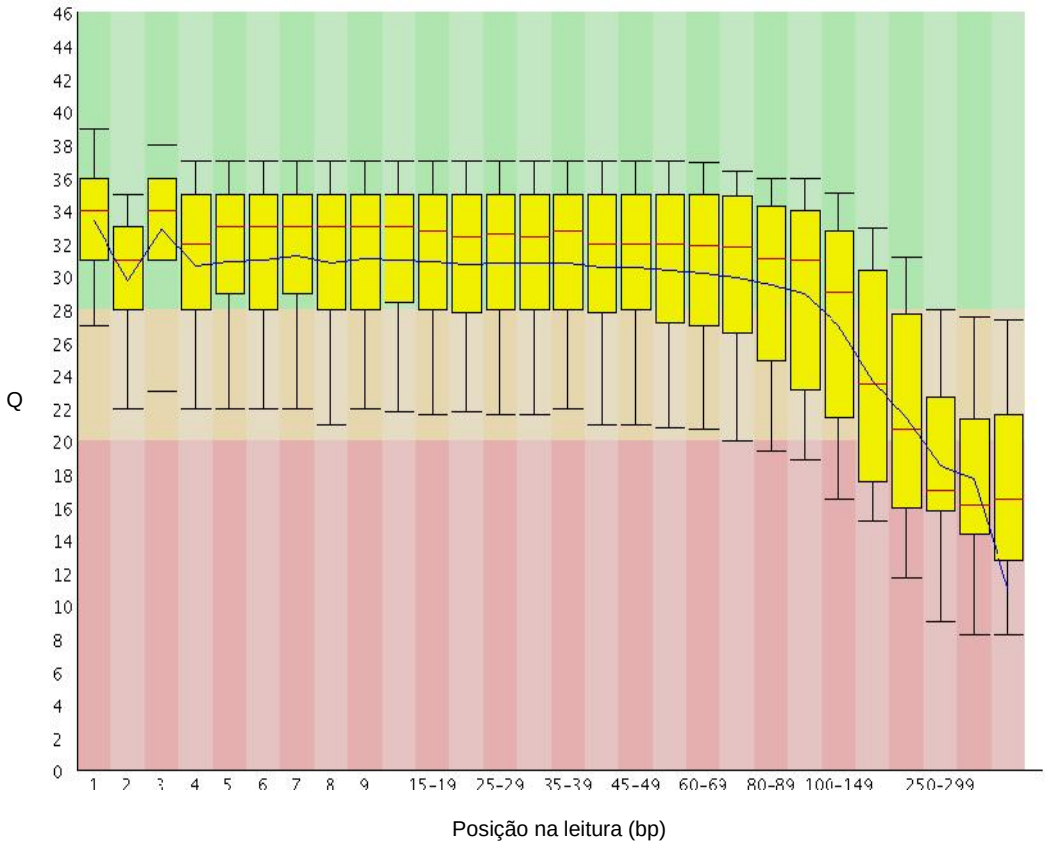

Figura 6. Diagrama de extremos e quartis, produzido pela ferramenta "FastQC v. 0.10.0" utilizado para analisar a qualidade das bases lidas. A qualidade, $Q = -10\ Log_{10}\ P$ em que P = probabilidade da base estar incorrecta. Para cada posição, a caixa amarela representa a amplitude inter-quartil (25-75%) e as linhas verticais que dela partem unem-na aos extremos inferior e superior. A linha vermelha representa a mediana e a linha azul representa a qualidade média.

Nas 99.786 sequências não redundantes, o software "MsatCommander v. 0.8.2" encontrou um total de 1185 loci microssatélites, com uma distância média calculada entre cada um de 10,21 kb (Quadro 3). Entre os microssatélites, os dinucleótidos foram a forma mais representada (~ 95 % do total).

Quadro 3. Distribuição de frequências dos tipos de microssatélites identificados nas 99.876 sequências não redundantes de *A. unedo*.

Repetição	Número de loci identificados	Percentagem (%)	Frequência (%)	Distância média entre SSRs (kb)
Dinucleótidos	1131	95.44	1.133	10.70
Trinucleótidos	48	4.05	0.048	252.17
Tetranucleótidos	4	0.34	0.004	3026.02
Pentanucleótidos	0	0.00	0.000	-
Hexanucleótidos	2	0.17	0.002	6052.04
Total de SSRs	1185	100	1188	10.21

Nota: Frequência = número de SSRs / número total de sequências não redundantes; Distância média entre SSRs = comprimento total das sequências não redundantes / número total de SSRs.

O motivo AG foi o mais frequente, representando aproximadamente 72 % de todos os microssatélites, e o motivo CG foi o menos frequente, à semelhança do descrito no trabalho de Zhu *et al*. (2011) acerca da identificação de SSRs em *Vaccinium macrocarpon* Ait.. O motivo CG, é além do mais, o dinucleótido mais raro em todos os genomas estudados até hoje. Ainda no trabalho anterior, os autores constataram que o motivo trinucleótido AAG foi o mais comum e CGG o mais raro, como também se verificou em *A. unedo* (Figura 7). Cavagnaro *et al*. (2010) também constataram esta frequência relativa dos motivos trinucleótidos em *Cucumis sativus* L. e nas dicotiledóneas em geral, referindo ainda que o oposto ocorre nas plantas monocotiledóneas, como arroz e sorgo, onde o motivo CGG é de longe o mais abundante.

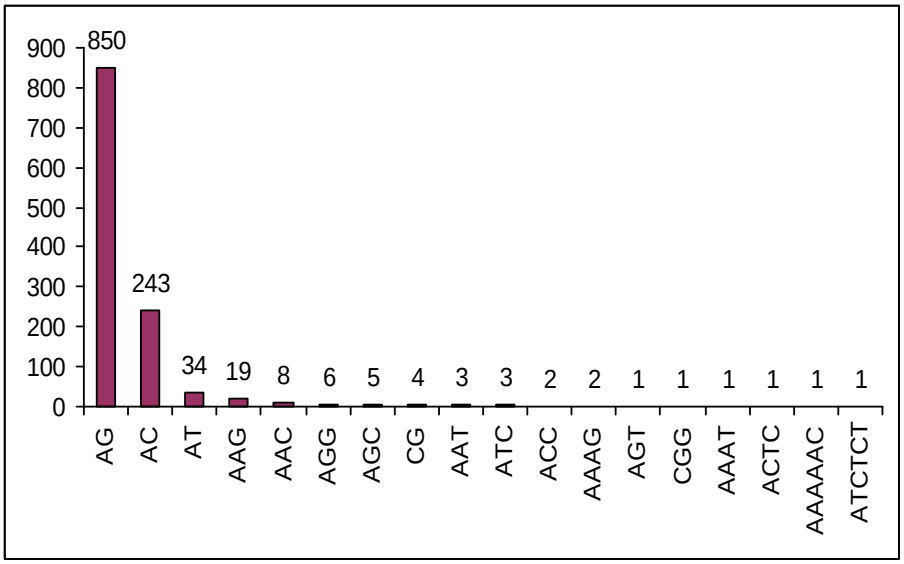

Figura 7. Distribuição de frequências dos motivos microssatélites identificados em *A. unedo*.

Certos motivos microssatélites, são comuns no reino *Plantae*, sendo alguns muito específicos de uma certa família ou género. No caso do medronheiro, como não existe nenhuma outra espécie do género *Arbutus* sequenciada, o termo de comparação mais próximo poderá vir do género *Vaccinium* e como acima visto os resultados não foram muito diferentes em termos de distribuição de motivos.

Os 1185 loci microssatélites, foram identificados em 1085 sequências que foram submetidas nas bases de dados do NCBI (http://www.ncbi.nlm.nih.gov/nuccore/?term=arbutus%20unedo%20fazenda). Para avaliar o uso potencial dos microssatélites identificados, foram desenhados primers para 18 loci (Quadro 4). Todos os marcadores microssatélites foram amplificados com sucesso e 3 revelaram-se polimórficos (Figura 8).

Quadro 4. Características dos dezoito loci microssatélites amplificados.

Locus / Nº acessão Genbank	Sequência do primer (5' - 3')	Ta (ºC)	Motivo da repetição	Tamanho esperado (bp)	Polimórfico
AU1427/ KF023647	F: GAAATATAAGCCCAAAATCAGC R: GCAGAAACCTATGCTCATC	60	$(AG)_7$	93	Sim
AU10205/ KF023705	F: CAAACTGTGGCAGTATGAG R: TCTAATTCTTCAGCATGATATGG	57	$(TA)_6$	96	Não
AU13867/ KF023732	F: AATTTGAATAATATCAAGGGGAATC R: GGGAAAGAAATCCAACTTTTTC	57	$(AG)_6$	117	Não
AU18473/ KF023768	F: CAATCGGATAAAAAATTAATACCTC R: GGTTCTTTGACGAGTTACTAT	57	$(CT)_6$	99	Não
AU18938/ KF023771	F: AATTTTAGGAGAAAGTGGAAAG R: ACGATACGAACAACAATAATAAG	57	$(CT)_8$	120	Não
AU25325/ KF023821	F: GGATAACGGATTCTTCCTAG R: CCATTATACTTTCATCTTGAAAAAGG	57	$(AG)_6$	106	Não
AU32030/ KF023875	F: ATTTGAGGTATCCACAACATG R: GCAGTATTCGCCATCTAAG	57	$(GT)_6$	124	Não
AU39217/ KF023934	F: TACCTTTGAGAGCTTTCTAAG R: TACGAGTCTTCTCACAATCT	57	$(AG)_6$	117	Não
AU58035/ KF024102	F: ACTTACGGATAAGGCATTC R: TACTCTTGTCGAATAGCATTC	57	$(AG)_7$	117	Não
AU65100/ KF024173	F: TAAGAACGTATCAATGGGC R: TTCAAGATGGTGTTCCTAATAC	57	$(GT)_6$	115	Não
AU69656/ KF024218	F: ATTGAGCGACAGAACTAGT R: CTGTAACTCATGCACGAA	58	$(AG)_9$	93	Sim
AU71512/ KF024233	F: AATTGAAATAGGTAGCTCAAGC R: GAAAGGACGCAATTGTTG	57	$(AG)_{12}$	125	Não
AU81604/ KF024305	F: AATTTGATCGAACTTCACAC R: TACTTATCCAAACTCTGAAGG	58	$(AG)_8$	115	Sim
AU93953/ KF024411	F: TGGTAAACAGTATTAAGGACAG R: GTAGGTTTTGCCCTACAG	57	$(AG)_6$	114	Não
AU95244/ KF024426	F: GAATCAAAGGTTTGGAGTTG R: GTCAGATCTTCCGGTCA	57	$(AG)_{11}$	106	Não

AU97280/ KF024448	F: CCTGGAGTTACTATTAAGCC R: TACAGAGATCAGATAGCTAGTG	57	$(GT)_6$	101	Não
AU98994/ KF024458	F: TTCTAGGAAATTGTTGAGGC R: GATAGTAAGCCCTTTGTATCC	57	$(GT)_8$	120	Não
AU118386/ KF024616	F: GCGAAACAACGCAGATC R: CAGAGAGTGGTTGTAGAGAG	57	$(CT)_6$	103	Não

Ta = Temperatura de "annealing"

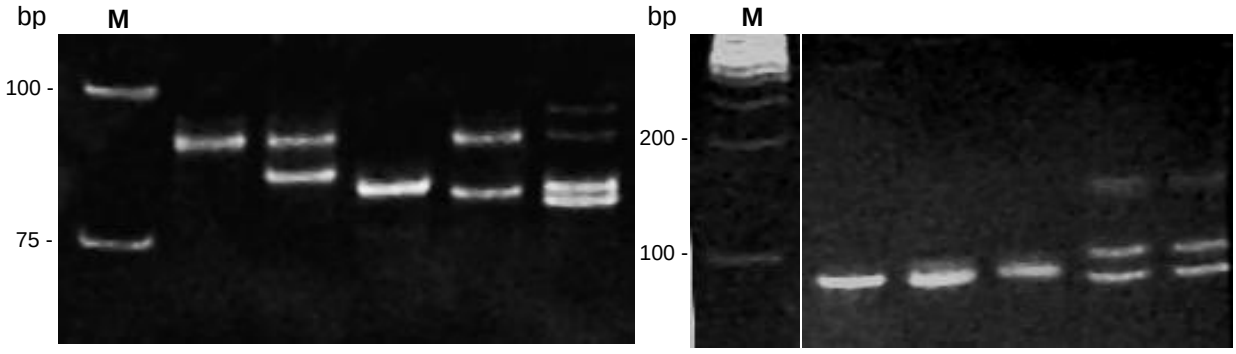

Figura 8. Amplificação de dois loci microssatélite polimórficos em cinco das amostras de Odiáxere. Gel de poliacrilamida (10 %). A: M – Marcador de peso molecular "DNA Ladder Ultra Low Range" (Fermentas), Locus AU 1427; B: M – Marcador de peso molecular "DNA Ladder Mix" (Fermentas), Locus AU 69656.

Identificação de SNPs

Para a identificação de SNPs optou-se por fazer um alinhamento *de novo* partindo dos "raw data" no "GS *De Novo* Assembler" do qual resultaram 1607 "contigs" com tamanhos que variam entre 102-917 bp, com um comprimento médio de 690 bp (Quadro 5).

Quadro 5. Resultados do alinhamento *de novo* para identificação de SNPs.

Número total de sequências > 50 bp	188.208
Número de sequências únicas	130.784
Número de sequências alinhadas	45.216
Número de sequências atípicas (quiméricas, artefactos, baixa qualidade)	12.208
Número total de contigs	1607
Tamanho médio dos contigs (bp)	690

A exploração manual no software “Tablet v.1.13.05.17” dos dez maiores “contigs” (o menor dos quais com 422 bp) resultou na identificação de vinte e cinco SNPs. (Anexo). Destes, doze foram A/G, onze C/T e apenas dois A/C.

Foi testado para marcador CAPS o “contig” que possuía um sítio de corte com uma enzima de restrição em stock no laboratório (Quadro 6).

Quadro 6. Características da sequência testada para marcador CAPS.

Nº do Contig	Sequência do primer (5’ - 3’)	Ta (ºC)	Tamanho do produto amplificado (bp)	Enzima de restrição utilizada	Ponto de corte (5’ - 3’)
1	F: TTACTAACAATGCAGGGC R: CAGCTGGTTGCTTATTAAG	58	380	DraI	TTTAAA

Ta = Temperatura de “annealing”

Para além do ponto de corte criado pelo SNP (C/T), existe também um outro ponto de corte, em ambos os alelos, reconhecido pela mesma enzima (Figura 9). Os tamanhos esperados dos fragmentos são para o alelo 1: 90 e 290 bp e para o alelo 2: 82, 90 e 208 bp (Figura 9, Figura 10). Este CAPS foi analisado com sucesso nas amostras de Gambelas e Odiáxere.

1

TTACTAACAATGCAGGGCAAGTTGAGATGGCTAGGATGAGACTTTTTCCATTCTCTCTAAAA
GACAAAGCCAAAGTTGGCTTACCAC**TTTAAA**GCCAATACTGTGCCTATTGGGTTTCGATGC
AAGCTGAATATCCTTAAACGTCTTTTCCCGATGCATCGAACCACTCGC**TCTAAA**GAAACAA
ATTCAAAACTTTTCTGAGAGACCCAATGAAGACTTTTTGTGAATGTTGGGAACGGTTTAAGG
AATATCTGTCAGCCATTCCACATACGGTTATGATGACTGGCAATTGGTTGCTTTCTTTTACG
AAAATATCTCTGCACGCAATCGCCAGTTTATAAACATGATGTGTAATGCCGATTTCCTTAATA
AGCAACCAGCTG

2

TTACTAACAATGCAGGGCAAGTTGAGATGGCTAGGATGAGACTTTTTCCATTCTCTCTAAAA
GACAAAGCCAAAGTTGGCTTACCAC**TTTAAA**GCCAATACTGTGCCTATTGGGTTTCGATGC
AAGCTGAATATCCTTAAACGTCTTTTCCCGATGCATCGAACCACTCGC**TTTAAA**GAAACAA
ATTCAAAACTTTTCTGAGAGACCCAATGAAGACTTTTTGTGAATGTTGGGAACGGTTTAAGG
AATATCTGTCAGCCATTCCACATACGGTTATGATGACTGGCAATTGGTTGCTTTCTTTTACG
AAAATATCTCTGCACGCAATCGCCAGTTTATAAACATGATGTGTAATGCCGATTTCCTTAATA
AGCAACCAGCTG

Figura 9. Produtos amplificados do “contig” 1 mostrando os pontos de corte reconhecidos pela enzima DraI (TTTAAA) para os dois alelos. No alelo 1, a presença de uma citosina impede o corte com a enzima de restrição. No alelo 2, a presença de uma timina permite o corte enzimático.

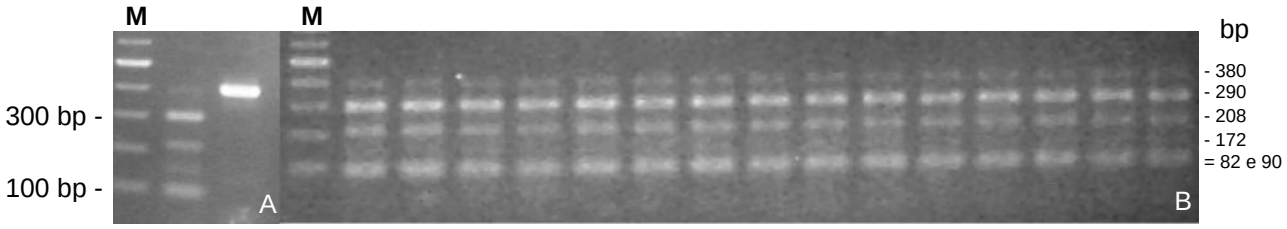

Figura 10. Gel de agarose (2 %). M - Marcador de peso molecular “DNA Ladder Mix” (Fermentas). “Contig” 1 - A: Produto de PCR cortado e não cortado da amostra Gb. B: Produtos de PCR cortados nas quinze amostras Od, todas heterozigóticas.

No painel de dezasseis indivíduos (Gambelas e Odiáxere) não foi possível identificar polimorfismos por marcadores CAPS, sendo no entanto possível concluir que os indivíduos analisados são todos heterozigóticos neste *locus*, uma vez que em todos se obtiveram os fragmentos esperados para os dois alelos. Para além dos fragmentos esperados observam-se vestígios do produto original (não cortado) e um fragmento de 172 bp resultante de uma digestão parcial no alelo 2: fragmentos obtidos por digestão no 2º ponto de corte e não no 1º (Figura 9, Figura 10).

Conclusões

O protocolo de extração de DNA a partir de núcleos mostrou-se rápido e eficiente. O tampão de extração utlilizado é mais simples e o DNA obtido aparentemente de melhor qualidade do que até ao momento obtido por outros autores (Sá *et al*., 2011).

A sequenciação massiva paralela de DNA genómico de *A. unedo* originou 198.856 sequências, disponibilizadas na base de dados "Sequence Read Archive" (SRA) do NCBI, e possibilitou a identificação de 1085 sequências contendo microssatélites, as primeiras publicadas para a espécie.

Dos 18 loci microssatélites para os quais se desenhou primers, foram identificados 3 marcadores polimórficos num painel de dezasseis amostras, uma de Gambelas (usada para a sequenciação) e quinze da empresa Cortevelada, Odiáxere. Os restantes loci apresentaram-se monomórficos em todas as amostras, o que sugere um baixo nível de diversidade genética nesta população.

Com o objectivo de identificar mais polimorfismos, explorou-se outro tipo de marcadores. Foram identificadas 10 sequências contendo 25 SNPs. Para uma destas sequências foi desenvolvido um marcador CAPS, no entanto todas as 16 plantas analisadas com este marcador apresentaram um padrão heterozigótico e monomórfico.

O grau de variabilidade genética revelado neste trabalho para a população de Odiáxere é inferior ao que seria expectável numa população obtida por via seminal numa espécie de polinização cruzada, como tem sido considerado o medronheiro. Os resultados obtidos levantam por isso dúvidas quanto a esta consideração e apesar de Hagerup (1957) ter levantado esta questão após constatar o vingamento de frutos em plantas de medronheiro cultivadas em estufa cujas flores estavam isoladas com sacos de plástico, não foram publicados desde então novos trabalhos sobre este assunto.

Referências bibliográficas

Alarcão-e-Silva, M.L.C.M.M.; Leitão, A.E.B.; Azinheira, H.G.; Leitão, M.C.A., **2001**. The Arbutus Berry: Studies on its Color and Chemical Characteristics at Two Mature Stages. J. Food Compos. Anal 14, 27-35.

Armour, J.A.; Neumann, R.; Gobert, S.; Jeffreys, A.J., **1994**. Isolation of human simple repeat loci by hybridization selection. *Human Molecular Genetics*, 3, 599-565.

Bassil, N.V.; Bunch, T.; Nyberg, A.; Hummer, K.; Zee, F.T., **2010**. Microsatellite Markers Distinguish Hawaiian Ohelo from Other *Vaccinium* L. Section *Myrtillus* Species. Acta Hort. 859, ISHS 2010.

Batley, J.; Barker, G.; O'Sullivan, H.; Edwards, K.J.; Edwards, D., **2003**. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol., 132(1):84–91.

Bingre, P.; Aguiar, C.; Espírito-Santo, D.; Arsénio, P.; Monteiro-Henriques, T., [Coord.s Cient.], **2007**. Guia de campo – As árvores e os arbustos de Portugal continental. p 331 in vol. IX dea Sande Silva, J. [Coord. Ed.] (2007): Coleção Árvores e Florestas de Portugal. Jornal Público/ Fundação Luso-Americana para o Desenvolvimento/ Liga para a protecção da Natureza. Lisboa. 9 vols.

Blankenberg, D.; Gordon, A.; Von Kuster, G.; Coraor, N.; Taylor, J.; Nekrutenko, A.; e a "equipa Galaxy", **2010b**. Manipulation of FASTQ data with Galaxy. Bioinformatics, 26(14):1783-5.

Blankenberg, D.; Taylor, J.; Schenck, I.; He, J.; Zhang, Y.; Ghent, M.; Veeraraghavan, N.; Albert, I.; Miller, W.; Makova, K.D.; Hardison, R.C.; Nekrutenko, A., **2007**. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. Genome Res. 17(6):960-4.

Blankenberg, D.; Von Kuster, G.; Coraor, N.; Ananda, G.; Lazarus, R.; Mangan, M.; Nekrutenko, A.; Taylor, J., **2010a**. Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*. Chapter 19:Unit 19.10.1-21.

Boches, P.; Bassil, N.V.; Rowland, L., **2006**. Genetic diversity in the highbush blueberry evaluated with microsatellite markers. J. Amer. Soc. Hort. Sci 131(5):674-686.

Boches, P.S.; Bassil, N.V.; Rowland, L.J.; **2005**. Microsatellite markers for *Vaccinium* from EST and genomic library. Mol. Ecol. 5, 657-660.

Brookes, A.J., **1999**. The essence of SNPs. Gene, Vol. 234 (2), 177-186.

Cavagnaro, P.F.; Senalik, D.A.; Yang, L.; Simon, P.W.; Harkins, T.T.; Kodira, C.D.; Huang, S.; Weng, Y., **2010**. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). BMC Genomics 2010, 11:569.

Celikel, G.; Demirsoy, L.; Dermirsoy, H., **2008**. The strawberry tree (*Arbutus unedo* L.) selection in Turkey. Scientia Horticulturae 118, 115–119.

Doyle, J.J. e Doyle, J.L., **1987**. A rapid DNA isolation method for small quantities of fresh tissues. Phytochem. Bull. Bot. Soc. Amer., 19, 11-15.

Ender, A.; Schwenk, K.; Stadler, T.; Streit, B.; Schierwater, B., **1996**. RAPD identification of microsatellites in *Daphnia*. *Molecular Ecology*, 5, 437-441.

Ertekin, M.; Kirdar, E., **2010**. Breaking seed dormancy of strawberry tree (*Arbutus unedo*). *Int. J. Agric. Biol.,* 12: 57–60.

Faircloth, B.C. **2008**. MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. Molecular Ecology Resources 8:92-94.

Fang, G.; Hammar, S.; Grumet, R., **1992**. A quick and inexpensive method for removing polysaccharides from plant genomic DNA. BioTechniques, Vol.13, No. 1.

Fiorentino, A.; Castaldi, S.; D'Abrosca, B.; Natale, A.; Carfora, A.; Messere, A.; Mónaco, P., **2007**. Polyphenols from the hydroalcoholic extract of *Arbutus unedo* living in a monospecific Mediterranean woodland. *Biochemical Systematics and Ecology*, 35, 809–811.

Georgi, L.; Herai, R.H.; Vidal, R.; Carazzolle, M.F.; Pereira, G.G.; Polashock, J.; Vorsa, N., **2012**. Cranberry microsatellite marker development from assembled next-generation genomic sequence. Mol Breeding, 30:227–237.

Giardine, B.; Riemer, C.; Hardison, R.C.; Burhans, R.; Elnitski, L.; Shah, P.; Zhang, Y.; Blankenberg, D.; Albert, I.; Taylor, J.; Miller, W.; Kent, W.J.; Nekrutenko, A., **2005** Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451-5.

Goecks, J.; Nekrutenko, A.; Taylor, J.; e "a equipa Galaxy" **2010**. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11(8):R86.

Gomes, F.; Costa, R.; Ribeiro, M.M.; Figueiredo, E.; Canhoto, J.M., **2012**. Analysis of genetic relationship among *Arbutus unedo* L. genotypes using RAPD and SSR markers. Journal of Forestry Research DOI 10.1007/s11676-012-0302-0.

Hagerup, O., **1957**. Wind autogamy in Arbutus. Bulletin du Jardin botanique de l'État a Bruxelles, vol. 27, Fasc. 1 (Mar. 31), pp. 41-47.

Hirai, M.; Yoshimura, S.; Ohsako, T.; Kubo, N., **2010**. Genetic diversity and phylogenetic relationships of the endangered species *Vaccinium sieboldii* and *Vaccinium ciliatum* (Ericaceae) Plant Syst Evol, 287:75–84.

Jennings, T.N.; Knaus, B.J.; Kolpak, S.; Cronn, R., **2011**. Microsatellite primers for the pacific northwest endemic conifer *Chamaecyparis lawsoniana* (Cupressaceae). American Journal of Botany, e323–e325.

Kalendar, R., **2007**. FastPCR: a PCR primer and probe design and repeat sequence searching software with additional tools for the manipulation and analysis of DNA and protein. (www.biocenter.helsinki.fi/bi/programs/fastpcr.htm).

Karagyozov, L.; Kalcheva, I.D.; Chapman, V.M., **1993**. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. *Nucleic Acids Research*, 21, 3911-3912.

Kijas, J.M.; Fowler, J.C.; Garbett, C.A.; Thomas, M.R., **1994**. Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *Biotechniques*, 16, 656-662.

Kivçak, B. e Mert, T., **2001**. Quantitative determination of [alpha]-tocopherol in *Arbutus unedo* by TLC-densitometry and colorimetry. Fitoterapia 72, 656-661.

Konieczny, A. e Ausubel, F.M., **1993**. A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant J. 4, 403-410.

Li, W., e Godzik, A., **2006**. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22: 1658-1659.

Li, W.; Jaroszewski, L.; Godzik, A., **2001**. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 17: 282-283.

Li, W.; Jaroszewski, L.; Godzik, A., **2002**. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics, 18: 77-82.

Litt, M. e Luty, J.A. **1989**. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics*, 44, 397-401.

Lopes, L.; Sá, O.; Pereira, J.A.; Baptista, P., **2012**. Genetic diversity of Portuguese *Arbutus unedo* L. populations using leaf traits and molecular markers: An approach for conservation purposes. Scientia Horticulturae 142, 57–67.

Lynch, M., **1990**. The similarity index and DNA fingerprinting. Mol. Biol. Evol. 7, 478-484.

Males, Z.; Plazibat, M; Vundac, V.B.; Zuntar, I., **2006**. Qualitative and quantitative analysis of flavonoids of the strawberry tree – *Arbutus unedo* L. (*Ericaceae*). *Acta Pharm. 56,* 245–250.

Mereti, M.; Grigoriadou, K.; Nanos, G.D., **2002**. Micropropagation of the strawberry tree, *Arbutus unedo* L.. Scientia Horticulturae, 93, 143–148.

Milne, I.; Bayer, M.; Cardle, L.; Shaw, P.; Stephen, G.; Wright, F.; Marshall, D.; **2010**. Tablet – next generation sequence assembly visualization. Bioinformatics 26(3), 401-402.

Mulas, M.; Cani, M.R.; Brigaglia, N.; Deidda, P., **1998**. Varietal selection in wild populations for the cultivation of myrtle and strawberry tree in Sardinia. *Rivista di Frutticoltura e di Ortofloricoltura,* 60 (3), 45–50.

Oliveira, I.; Coelho, V.; Baltasar, R.; Pereira, J.A.; Baptista, P.; **2009**. Scavenging capacity of strawberry tree (*Arbutus unedo* L.) leaves on free radicals. Food and Chemical Toxicology 47, 1507-1511.

Pawlowska, A.M.; Marinella, D.L.; Braca, A., **2006**. Phenolics of *Arbutus unedo* L. (Ericaceae) fruits: identification of anthocyanins and gallic acid derivatives. J. Agric. Food Chem. 54, 10234-10238.

Pedro, J.G., **1994**. Carta da distribuição de figueira e medronheiro - Notícia Explicativa II.6. D.G.A., Ministério do Ambiente e Recursos Naturais, Lisboa.

Peterson, D.G.; Boehm, K.S.; Stack, S.M., **1997**. Isolation of Milligram Quantities of Nuclear DNA from Tomato *(Lycopersicon esculentum),* A Plant Containing High Levels of Polyphenolic Compounds. Plant Mol. Biol. Reptr. 15(2):148-153.

Rassmann, K.; Schlotterer, C.; Tautz, D., **1991**. Isolation of simple-sequence loci for use in polymerase chain reaction-based DNA fingerprinting. *Electrophoresis*, 12, 113-118.

Rozen, S. e Skaletsky, H.J., **2000**. Primer3 on the WWW for general users and for biologist programmers. In: Bioinformatics methods and protocols (eds. Krawetz, S., e Misener, S.), 365–386. Humana Press, Totowa, New Jersey, USA.

Sá, O.; Pereira, J.A.; Baptista, P., **2011**. Optimization of DNA extraction for RAPD and ISSR analysis of *Arbutus unedo* L. leaves. *Int. J. Mol. Sci.* 2011, *12*, 4156-4164.

Shendure, J. e Ji, H., **2008**. Next-generation DNA sequencing. Nat. Biotechnol. Vol. 26, No. 10.

Sulusoglu, M., **2012**. Development of a Rooted Cutting Propagation Method for Selected *Arbutus unedo* L. Types and Seasonal Variation in Rooting Capacity. Journal of Agricultural Science, Vol. 4, No. 11.

Sulusoglu, M.; Cavusoglu, A.; Erkal, S., **2011**. *Arbutus unedo* L. (Strawberry tree) selection in Turkey Samanli mountain locations. J. Med. Plant. Res. Vol. 5(15), pp. 3545-3551.

Takrouni, M.M.; Boussaid, M., **2010**. Genetic diversity and population's structure in tunisian strawberry tree (*Arbutus unedo* L.). Scientia Horticulturae, 126, 330–337.

Tautz, D. **1989**. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research*, 17, 6463-6471.

Tawfik, D.S.; Griffiths, A.D., **1998**. Man-made cell-like compartments for molecular evolution. Nat. Biotechnol., 16, 652–656.

Tilki, F., **2004**. Improvement in seed germination of *Arbutus unedo* L. Pak. J. Biol. Sci., 7(10): 1640-1642.

Torres, J.A.; Valle, F.; Pinto, C.; Garcia-Fuentes, A.; Salazar, C.; Cano, E., **2002**. *Arbutus unedo* communities in southern Iberian Peninsula mountains. Plant Ecol. 160, 207-223.

Wang, N.; Qin, Z.C.; Yang, J.B.; Zhang, J.L., **2010**. Development and characterization of 15 microsatellite *loci* for *Rhododendron delavayi* Franch. (Ericaceae). Hortscience 45, 457-459.

Weber, J.L. e May, P.E., **1989**. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics*, 44, 388-396

Yang, H.; Tao, Y.; Zheng, Z.; Shao, D.; Li, Z.; Sweetingham, M.W.; Buirchell, B.J.; Li, C., **2013**. Rapid development of molecular markers by next-generation sequencing linked to a gene conferring phomopsis stem blight disease resistance for marker-assisted selection in lupin (*Lupinus angustifolius* L.) breeding. Theor Appl Genet 126:511–522.

Ying, H.; Beifang, N.; Ying, G.; Limin, F.; Weizhong, L., **2010**. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010(26): 680-682.

Zane, L.; Bargelloni, L.; Patarnello, T., **2002**. Strategies for microsatellite isolation: a review. Molecular Ecology 11, 1-16.

Zhu, H.; Senalik, D.; McCown, B.H.; Zeldin, E.L.; Speers, J.; Hyman, J.; Bassil, N.; Hummer, K.; Simon, P.W.; Zalapa, J.E., **2012**. Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (*Vaccinium macrocarpon* Ait.). Theor Appl Genet, 124:87–96.

Anexo

contig00001

TACTTTTACTAACAATGCAGGGCAAGTTGAGATGGCTAGGATGAGACTTTTTCCATTCTC
TCTAAAAGACAAAGCCAAAGTTGGCTTACCACTTTAAAGCCAATACTGTGCCTATTGGGT
TTCGATGCAAGCTGAATA**(T/C)**CTTAAACGTCTTTTCCCGATGCATCGAACCAC**(T/C)**GCT**(C/T)**TAAAGAAACAAATTCAAAACTTTTCTGAGAGACCCAATGA**(A/G)**ACTTTTTGTGAATG
TTGGGAACGGTTTAAGGAATATCTGTCAGCCATTCCACATACGGTTATGATGACTGGCA
ATTGGTTGCTTTCTTTTACGAAAATATCTCTGCACGCAATCGCCAGTTTATAAACATGATG
TGTAATGCCGATTTCCTTAATAAGCAACCAGCTGA**(T/C)**GCATTAGATTACTTGGATGAAT
TGGCAGCGACCAATAAATCATGGGATTACACTGATCCCCAAGAGCGCACTTATAAACAG
TCCTCATCTGTCCAAACTGCCCCTGGAAAATATGTATTGAAGGGAGAGGATGATCTCAA
TCGAAAGCTTGACCTACTGGCTAAAAAAATGGATAATCTGGAAATTCATAAGGTCAACGA
GGTCTCTTCTGTCCCTCGGCAAAGTGAGTTGTGCGTAATTTGCGAGACACCGGGGCAT
CCTACTCAAGAGTGCCCCACTATACCTGCTTTCAAAGAGGTATTACATGGTTCGGATCC
GGCTGAGGTGAATGCCCTAAATCAAGGACAAAAACCTAAGCAATGCTGCATACTATCCA
GGGTGGAGGAACCACCCGAACTTCAGCTGGAGAAATAATGAGTCTGTTGGGCCCACTG
CTGGCCCACCGGGATATTCGACTCAAGGGCCACAAAATTTCCAGAAATCAATTTCAAGG
CCCAAATCAAAATTTTGCTGCACCACAAGGCAGGTA

contig00002

AATTCAATATTCACATATATCATGTGCCTCATAGGATCATCTCATATCGTGTCGGCGTCC
GCACATATCATGTATCACATAACTATTTCTTAAGTTTATTTAAACCACATGGACCATAGGT
TCCATGACTCTCCCTGTGCATAATCTATCATTCTCATAACTCTCTCATAGATACATTCGTC
AGATTTGCTCAACAACGTCTCGAAAAGACAATCTATGTTCCATCATGACGCATCATACAA
TCTTCGTTCACAATATACCATAGAGTTCTACACAATTCCAACAATCAATGGAAAGTTAACA
ATTTGGTCACAAGAGCAATAAACACTTTTAGCGATACATGAATTTAAACAATGGAGGCAT
ATGAATAGCCCTCCGTCAACGTATCACTTGCAAATTACTGTGTGTCATCTCAACGTCAAA
GCATCTCGACGTATCTCGTGGCCATTACCCTTCATTACAA**(T/C)**GCCATGTATGACTTGT
ATTAA**(T/C)**GTCAATCAA**(C/T)**AGTGAAAACATTACTTCTTAATTAGCAGTTATCATGACCC
ATTCTCGGTCCATATTTCACGCCTCATGTCCGTCGTCACTCCTAACTCTAACCCAATGGT
GGTTAGACAAGCCATGTCCTATACTAGCACAATCCCCAAGGGAACTAGAAAAGACTCTC
ATACGTTCCATTCAATTTCGATAATCTACACTGGCACCACACTCGAAGGTGGTCCAATAA
GATATGTTATCTTACTAACACGATTCTCACGGAAAAGTAGCCAACTTTCATTAACCATTAC
ATTGTTTCGTATATTCCATTCAAACTCATTTGTCACGGACGGCGCCTCACCCGAAAG

contig00006

AAATTTAGCATAAGATGGTATCTGTTTAATAGCATCGAGCAGGGGAATATTCACCTTAAC
TGTTTAAACAATTCATAGATTTCAGCAGTGTGACTGGGTTTGGGCTGGGCCCGCAATCT
ATGAGGAAATGGTGCTGGAACAGG**(G/A)**CAGGTGGAGACCACAGAGTCTCTCTCAGGT
GTGCTTACGGGTGGTGGTGCGGAAGTAGTAGGAGTGTCCCCCGAACTAGAAGTAGGAG
AGGGGGCATGAGGAGGAACAGCTTTGGGATGGATAGTCTTATCGACTTCCTTCCCATTC
CTCAAAATGGTGATGGATTTGGCTTGCTCCAGATTTGGAGGTGC**(A/G)**GAGCTACTCAC
AAAGTGCTGGCCTTGAGGATTGGCCTGTGGTTGGGCAGGAAACCTGCCCTTCTCTTGTT
GTAAGGTCCCCAAGGTCGTGGTGATCTTAGCCATTTGAGAACGAATGTCTCCGATGGCT
TGTGTATTCTGCTCATTTATGTTTG**(A/C)**TTGCCCTTCCATAAATTTAGACAATAGATCCTC
AAGATTCCGCTTTGGAGGCGGTTGGTAAGTATTGGGACCCACAGAAGGTCCTTGAGGA
GCATAAGACT

contig00008

TGGCCAATGAGGCATATTATCTTACTAGCACCATTCCCTCGGAAGCTAGACGGTTCTTTC
ATAACTAGCACCACTCCTTCAAGGAGGCTAGAATTGATTCGATATCACTAGCACCTCTCC
TTTCGGAGGCTAGATCATAGTTCATCTCACTAGCACTTCTCCTCT**(C/T)**AGAGGCTAGAC
CATATTCATTTCACAGTCGAGATTTCGTTATTTGATAGGTTCC**(A/G)**TTCAAGTTCAAGGA
TCAATAACGGCACCTCACCCATAGGTGGCCTAATACAAATAAATATCTTACTAGCACAAT
TCCCTCGGGAAGCTAGTCAAATTTCATAACTAGCACCACTCCTATAAGGAGGCTAGCAT
TATTTCGTATTTTACTAGCACCGCTCCCATTGGAGGCTAGACCATTTGTTATGATCCCTA
GCACATCTCCTCACGGAGGCTAGACCACTACGCCTTGTGTCAATTTCGTATAAATCACA
AGTCTACACGGTGGATATTTCACAATTTCTAATCTCATGTCACCCGAGCCATTTACCGTC
ACTTATAAAACTTACCTATATCCAAGACAATACTCATATATCATACTCAACCCACAACCAA
CTTACCAAACTACTTATGAACAAAATGATTCAACCGAACACATAGGAAAAACTCAATCAA
AGGAGAAGGAGTA

contig00009

AATTGAGGAAGTAGGGATGCAACTTCATGTCTTCGAGCGCACAAGTCATTGCGAGAGCA
AACTGTGAGAATTGGAATCGAGGCTCGCTTTGGATAATAGACTTAGCTATACTCCAGGC
TCACGTGTTCGTTAGGGTTCATGTTGATTGCGTAAGAATGAATGGTGACTCATGGTTTG
GAATCATTCGTTTTCCATAGATCAAACTACAAATCAAGCATAGGGTAGAATCGCATTCGA
ATGGGAAATTTCTTTTCTTTTCTTCTTTAGTTTTTGGTATGATGCGTTGCTAGGGACTAGC
AACGTTCAAGTTGGGGGGTGTGTTGAGGGTACTAAATGATAGACTATTATAGCTCAATA
CATCCACTTTAACATCATCTGCGTATGGAAAATCATTGCGTTATTTAATATTGCACGCGTA
GATGTTGATGCTTCCTAACAGGTTTTCAAGAGCTAAATACAT**(A/G)**CCAGGGGGCTCGAG
CACGAGGCAAAAGGCCGGGACCAAATCAACGTCATTAAAAACCTAAGAGTAGGTCCGG
GGAATGAGAGAAGAGAAGCCGAAAAGGAAATTTGGAGAAGCCTGGGAAAATGAGTCAG
AAGGGTTCTCTATCGGTAGAGACCAATT

contig00010

AATTTATTTATTCTTTTTTTAACTGCAATTCTTCATCCCGACAGCTACACAAACACTGACG
GTCCAAATTGTAGAATATGATAAATGTCACTTCGATCGCAATAATCTGTCGTGAAAAGAT
AACGTGCACCGCCCGTTAAATTTTCTTTCCAAATACACTTGGTATGGGCTACCTATGTCA
CCCGTGTATTTATTGAAATTCCGAGATTATTGGGATTTTTAACTAG**(G/A)**TGATA**(C/T)**GGA
CAGCGTGTCAGCGTGAACTAACAGCACTATTC**(A/G)**TATTACTTAGCTATGAAGTAACGA
ATTTTTGCTTTTATTAATTTTTACGTGAAACAAAAAGAAAATAGCGGTAACTGTTACGGGG
AGACGGGAAAGGGGTTGAAAATAATAAAAAAGGTTAGGAATTTTGGGAAAGGAAAATAA
AAAGAAAAAATGTTACGGGATTCGGGAAAGGAAAATAAAAAGAAAAATGTTACGGGATT
CGGGAAAGGAAAAAGTTAGGAGTTAACTCTGACAGGTGCCGAGTCAATCGGGCGATAT
GTCGAGAGTGTAAATAAAAGAGAAAAGACAATAAAAAAAATGAAAAATGATATTCTGATA
TTCTAGAAATGCGGACTAGAAATT

contig00014

AAATTCGCCGGGATGCGTATGACAATGCCCACATTTACAAGTGTAAGGTCAAGGCCTTT
CACGATAAGAACATTGTTCGGAAGGT**(A/G)**TTCACCCCCGGCCAAAAAGTTTTGTTATAC
AACTCGAGGCTGCATATCTTTCCCGGTAAATTGCGATCCCGGTGGGGTGGCCCATACTT
TGTGAAAACCGTTTTCCCTCACGGAGCGGTCGAGATTGAGAATCCTAAAAATGGAACCA
CTTTCAAGGTGAATGGCCACCGTTTAAAACCTTTTCTTGAA**(T/C)**ACTTTGATCCCGAGGA
AATTGTAGAAAAGTTGAGTGACCCTATCTATAGGGAGTCCGAGGAGGACAATGAGGACC
CAAATTCTTCGGGTCAGTGAGCTTGAACCGAAGTCTAGGAGACTATAAATCCATGGAGT
TTGTGATACCTCCAGTGTTTGCCGTGCTCGAACGGCATCTGATGGTCTGGTATAGCCAC
CTTGTCAAGCTACGTATGTAGTA

contig00017
AATTGGGAGCATCCTAAGAGTGTGTTTGAGATTCGGAGCTTTTTGGGTTTGGCCGGATA
TTATCGACGATTTGTTCTTGACTTTTCGCGACTAGCGGCACCGATAACACAGTTGACAC
GTAAGGGGACACGGTTTGTGTGGGACGACAAGTGTGAGTCGGCATTTCAAGAGTTGAA
GAAACG**(G/A)**TTGACTAGCGC**(A/G)**CCTATCTTGATTGTGCCCGAGCGTGGGGTTGGCT
ACTCCGTTTATTGTGATGCTTCGCGGGAGGGGCTTGGTTGCGTGTTAATGCAAAATGGG
CGAGTTGTGGCATATGGGTCTCGACAACTTAAGACACATGAGCGTAACTATCCCACTCA
CGACTTAGAGTTGGCGGCTGTTGTCTTTGCATTGAAGAGTTGGAGGCATTATCTTTATG
GTGAGAAGTTTGAGGTGTTTTCGGATCACAAGAGCTTGAAGTATTTGTTCTCCCAAAAAG
AATT

contig00020
TACCTTGGTCACTAATGAGGACCTTAGGCATCCC**(G/A)**AACCTCGCAAAGATATGCTCCT
TGAGAAATTTCACAACAACCGAATGGTCATTGGTGCGAGTTGGGATAGCCTCTACCCAC
TTGGATACGTAGTCCACAGCAAGAAGATGAATAGATGCCCACAAGAGTTTGGGAATGGT
CCCATGAAGTCAATCTCCCAACAATCAAAGATTTCGACTATAAGGATGGGGGTGAGGGG
CATCATATTCCGGCGAGAAAGAGCTCCCAATTGTTGACACCTATCACATGATTTGCAATA
CTCGAAGGTGTCCTTAAAAAGATTTGGCCAATAGAAACCGCATTGTAAAACCTTGGCGG
CGGTTTTCTTAGAGGAGAAATGACCGCCGCATGCTTCAGTGTGGCAAAATGAAAGAATT
TCCTTTGTTCATCATTGGGCACACACCGCCGAATAATTT

contig00021
AATTCTCACGGTAAGATTGTGGAAAGTATTGGTTTTCAAACAAAGTCTCAAACTCCGCCC
AAGTGATGTTCTCAACATCCGGTTCCACAACAT**(A/G)**CCT**(C/A)**GCCGCCCTATCGGCTC
TCCGTGCATCCCTACGTGCCGACAAAATAGATTCCCACCACTCATTCGGCTCGCCGAAG
AGTTGACAAGCCATCAAACGGATCTTCGT**(T/C)**GTGTCACTAGTGATGTCCATGACATCA
AAACGCTTCCTCATTTCCGCAAGCCAATGGTTGGCATCAAGCGGATCCCCGGATTTCCC
ATCGAATTGTGGAGGGTGCAACCTACCAAATTTCTCGGCAATCTCCATGGCGCTAGGTG
TGTGGCTCACTCTTGGTTGGTTGATGGCGGTTGTTAAAGCGGTGACCAATTGAGCCAAA
AGTGGTCCTTGAGAA