UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Instituto de Biociências

Programa de Pós-Graduação em Genética e Biologia Molecular

Geração de recursos genômicos em *Eugenia uniflora* L. (Myrtaceae) usando tecnologias de sequenciamento de nova geração e ferramentas bioinformáticas

Frank Guzman Escudero

Porto Alegre

2014

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Instituto de Biociências

Programa de Pós-Graduação em Genética e Biologia Molecular

Geração de recursos genômicos em *Eugenia uniflora* L. (Myrtaceae) usando tecnologias de sequenciamento de nova geração e ferramentas bioinformáticas

Frank Guzman Escudero

Tese submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal do Rio Grande do sul como requisito parcial para obtenção do grau de Doutor em Ciências (Genética e Biologia Molecular).

Orientador: Prof. Dr. Rogerio Margis

Porto Alegre, Março de 2014

INSTITUIÇÕES E FONTES FINANCIADORAS

Este trabalho foi realizado no Laboratório de Genômica e Populações de Plantas, Centro de Biotecnologia e Departamento de Biofísica da Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil. O projeto foi subvencionado pelo CNPq, edital universal 2011. O doutorando obteve bolsa de estudos do CNPq (48 meses).

Á minha mãe Ines

AGRADECIMIENTOS

Á minha família, por sempre confiar e acreditar em mim.

Ao meu orientador, Rogerio, pela orientação, confiança, paciência e compreensão no desenvolvimento deste trabalho.

À professora, Marcia, pelo carinho, compreensão e conselhos.

Á minha relatora, Ana K., pelo tempo disponibilizado na revisão da minha tese.

Á minha banca, pela avaliação e crítica da minha tese.

Aos colegas do laboratório LGPP, Franceli, Priscila, Guilherme C., Ana C., Nureyev, Maria, Henrique, Caroline e Érika.

Aos ex-colegas do laboratório LGPP, Vanessa, Mauricio, Júlio, Guilherme L., Lorraine, Fernanda.

Aos amigos, Veronica, Melissa, Jennyfer, Corally, Johanna, Mauricio, Walter, Pablo, Frank M., John, Jimmy pelos momentos compartilhados durante esta viagem.

Á família PS e SdP pelos bons momentos nesta etapa da minha vida.

Ao Elmo pelo auxilio em cada momento.

Á CNPq, pela concessão da bolsa.

A todos aqueles que colaboraram na realização deste trabalho.

SUMARIO

ABREVIATURAS

AFLP - do inglês amplified fragment length polymorphism

C5 - hemiterpeno isopreno

C10 - monoterpenos

C15 - sesquiterpenos

C20 - diterpenos

C30 - triterpenos

C40 - pigmentos carotenoides

Cn - poliisopropenos

CoA - coenzima A (do inglês coenzyme A)

CPP - copalyl difosfato (do inglês copalyl diphosphate)

DCL-1 - do inglês endoribonuclease Dicer-like 1

DMAPP - difosfato de dimetilalilo (do inglês dimethylallyl diphosphate)

DNA - ácido desoxirribonucleico (do inglês deoxyribonucleic acid)

EBI - do inglês European Bioinformatics Institute

EC - Comissão de enzimas (do inglês Enzyme commission)

FDP - farnesil difosfato (do inglês farnesyl diphosphate)

FPP - farnesil pirofosfato (do inglês farnesyl pyrophosphate)

GO - do inglês Gene Ontology

GGPP - geranilgeranil pirofosfato (do inglês geranylgeranyl pyrophosphate)

GPP - geranil pirofosfato (do inglês geranyl pyrophosphate)

HYL1 - do inglês hyponastic leaves 1

IPP - isopentenil pirofosfato (do inglês isopentenyl pyrophosphate)

k-mer - subsequências de tamanho k

KEGG - do inglês Kyoto Encyclopedia of Genes and Genomes

MEP - metileritritolo fosfato (do inglês methylerythriol phosphate)

miRNA - microRNA

miRNA* - microRNA da fita passageira

mRNA - RNA mensageiro (do inglês messenger RNA)

Mpb - milhões de pares de bases (do inglês millions of base pairs)

NCBI - do inglês National Center for Biotechnology Information

NGS - sequenciamento de nova geração (do inglês next generation sequencing)

nt - nucleotídeos (do inglês nucleotides)

OSC - oxidoesqualeno ciclases (do inglês oxidosqualene cyclase)

pb - par de bases (do inglês base pair)

pre-miRNA - microRNA primário (do inglês primary microRNA)

pri-miRNA - microRNA precursor (do inglês precursor microRNA)

RAPD - do inglês random amplified polymorphic DNA

RISC - complexo de indução do silenciamento de RNA (do inglês RNA-induced silencing complex)

RNA - ácido ribonucléico (do inglês ribonucleic acid)

RNA-seq - sequenciamento do RNA (do inglês RNA Sequencing)

RT-PCR - do inglês reverse transcription polymerase chain reaction

SNP - do inglês single nucleotide polymorphism

TPS - terpeno sintase (do inglês terpene synthase)

RESUMO

Pitanga (*Eugenia uniflora* L.) é uma árvore arbustiva com frutas semelhantes à cereja, que cresce em diferentes tipos de vegetação e ecossistemas como consequência da sua alta capacidade de adaptação a diferentes condições de solo e clima. Esta espécie é de particular interesse econômico devido às suas propriedades medicinais, que são atribuídas aos metabólitos especializados presentes em suas folhas e frutos. Entre os metabólitos, os terpenóides são os mais abundantes dos óleos essenciais que são encontrados nas folhas. A diversidade de terpenos observada em Myrtaceae é determinada pela atividade de diferentes membros das famílias das terpeno sintases e oxidosqualeno ciclases. Por outro lado, os miRNAs são pequenos RNAs endógenos que desempenham papéis regulatórios essenciais no crescimento das plantas, desenvolvimento e resposta ao estresse. Por esta razão, estudos extensivos de miRNAs foram realizados em plantas modelos e outras plantas de importância econômica nos últimos anos. Portanto, o objetivo deste estudo é gerar recursos genômicos para *E. uniflora* utilizando sequenciamento de nova geração e ferramentas de bioinformática para a identificação de miRNAs conservados e novos, bem como os genes envolvidos na sínteses dos terpenóides.

No capítulo 3, bibliotecas de pequenos RNA e RNA-seq foram construídas a partir de folhas de pitanga para identificar miRNAs maduros e pre-miRNAs, respectivamente. De 14.489.131 leituras limpas de pequenos RNA, foram obtidos 1.852.722 sequências de miRNAs maduros que representam 45 famílias conservadas e que foram identificadas em outras espécies de plantas. Análises posteriores, usando *contigs* montados a partir do RNA-seq, permitiu a predição das estruturas secundárias de 42 pre-miRNAs: 25 conservados e 17 novos. Alvos potenciais foram previstos para os miRNAs maduros mais abundantes nos pre-miRNAs, identificados com base na homologia de sequências. Além disso, a expressão de 27 pre-miRNAs foi validada utilizando ensaios de RT-PCR em diferentes indivíduos de pitanga. Este estudo é o primeiro de identificação em grande escala de miRNAs e seus alvos potenciais de uma espécie da família Myrtaceae, e proporciona mais

informação sobre a conservação evolutiva dos vias regulatórias dos miRNAs em plantas com destaques para os miRNAs específicos da pitanga.

No capítulo 4, a biblioteca de RNA-seq sequenciada anteriormente foi utilizada para identificar os genes potencialmente envolvidos na via da biossíntese dos terpenos e diversidade dos terpenóides a partir da montagem *de novo* e a anotação do transcriptoma de *E. uniflora*. No total, foram identificados 72.742 unigenes com um comprimento médio de 1.048 pb. Destes, 43.631 e 36.289 unigenes foram anotadas com as bases de dados das proteínas não redundantes do NCBI e Swiss-Prot, respectivamente. A ontologia gênica categorizou as sequências em 53 grupos funcionais. A análise das vias metabólicas com KEGG revelou 8.625 unigenes designados a 141 vias metabólicas e 40 unigenes preditos como associados com a biossíntese de terpenóides. Por outro lado, foram identificados quatro genes putativos de terpeno sintases (TPS), de comprimento completo, envolvidos na biossíntese de monoterpenos e sesquiterpenos, e três genes putativos de oxidosqualeno ciclases (OSC), de comprimente completo, envolvidos na biossíntese de triterpenos. Além disso, a expressão destes genes foi validada em diferentes tecidos de *E. uniflora*. A futura caracterização bioquímica dos diferentes TPS e OSC descritos aqui determinará especificamente o tipo de terpenóide sintetizado em condições ambientais específicas.

Os dados produzidos neste estudo servirão como referência para estudos genéticos sobre os mecanismos moleculares que são responsáveis pela composição química dos óleos essenciais em *E. uniflora* e os aspectos fisiológicos da capacidade de adaptação desta espécie a diferentes habitats.

ABSTRACT

Pitanga (*Eugenia uniflora* L.) is a shrubby tree with edible cherry-like fruits that grows in different vegetation types and ecosystems as a consequence of its high adaptability to different soils and climate conditions. This species is of particular interest due to its medicinal properties that are attributed to specialized metabolites present in their leaves and fruits with potential pharmacological benefits. Among these metabolites, the terpenoids are the most abundant in the essential oils found in the leaves. The terpene diversity observed in Myrtaceae is determined by the activity of different members of the terpene synthase and oxidosqualene cyclase families. Furthermore, miRNAs are endogenous small RNAs that play essential regulatory roles in plant growth, development and stress response. For this reason, extensive studies of miRNAs have been performed in model plants and other plants of economic importance in the last years. Therefore, the aim of this study is to generate genomic resources for *E. uniflora* using next generation sequencing and bioinformatics tools to identify conserved and new miRNAs, and to identify the genes involved in the synthesis of terpenoids.

In chapter 3, small RNA and RNA-seq libraries were constructed from leaves to identify mature miRNAs and pre-miRNAs, respectively. From 14.489.131 small RNA clean reads we obtained 1.852.722 mature miRNA sequences, representing 45 conserved families that have been identified in other plant species. Further analysis using assembled contigs from RNA-seq allowed the prediction of secondary structures of 42 pre-miRNAs: 25 conserved and 17 novel. Potential targets were predicted for the most abundant mature miRNAs, which were identified in pre-miRNAs based on sequence homology. In addition, the expression of 27 identified pre-miRNAs was validated using RT-PCR assays in different individuals of pitanga. This study is the first large scale identification of miRNAs and their potential targets from a species of the Myrtaceae family. It provides information about the evolutionary conservation of the regulatory network of miRNAs in plants and highlights miRNA specific to pitanga.

In chapter 4, the previously sequenced RNA-seq library was *de novo* assembled followed by annotation of the *E. uniflora* transcriptome and used to

identify the genes potentially involved in the terpene biosynthesis pathway and terpenoid diversity. In total, we identified 72.742 unigenes with a mean length of 1.048 bp. Of these, 43.631 and 36.289 unigenes were annotated with the NCBI non-redundant protein and Swiss-Prot databases, respectively. The gene ontology categorized the sequences into 53 functional groups. A metabolic pathway analysis with KEGG revealed 8.625 unigenes assigned to 141 metabolic pathways and 40 unigenes predicted to be associated with the biosynthesis of terpenoids. Furthermore, we identified four putative full-length terpene synthase (TPS) genes involved in sesquiterpenes and monoterpenes biosynthesis, and three putative full-length oxidosqualene cyclase (OSC) genes involved in the triterpenes biosynthesis. In addition, expression of these genes was validated in different *E. uniflora* tissues. Future biochemical characterization of the different TPS and OSC described here will determine the type of terpenoid specifically synthesized in specific environmental conditions.

The data produced in this study will serve as a reference for genetic studies about the molecular mechanisms behind the chemical composition of the essential oils in *E. uniflora* and about the physiologic aspects of the capacity of adaptation of this specie to different habitats.

Capítulo 1

Introdução Geral

1. *Introdução*

1.1 A família Myrtaceae

A família Myrtaceae compreende gêneros de importância econômica e ecológica com uma distribuição concentrada nas regiões neotropical e australiana, com um total de 5.671 espécies agrupados em 132 gêneros no nível mundial (Govaerts *et al*., 2008). No Brasil, representa uma das principais famílias da flora, com 26 gêneros e aproximadamente 1.000 espécies (Souza e Lorenzi, 2005). A família está dividida em duas subfamílias: Myrtoideae, que apresenta frequentemente frutos em bagas, é de ampla ocorrência na América tropical e inclui os gêneros *Eugenia*, *Psidium* e *Syzygium*; e Leptospermoideae, com frutos secos e folhas alternas, ocorre na Australia e ilhas do Oceano Índico e inclui os gêneros *Eucalyptus*, *Melaleuca* e *Leptospermum* (Atchison, 1947).

Alguns gêneros desta família tem um alto número de espécies: *Syzygium*, *Eugenia* e *Eucalyptus* contém aproximadamente 1.500, 1.050 e 700 espécies, respectivamente (Brooker, 2000; Craven e Biffin, 2010). As espécies desta família se caracterizam pela sínteses de óleos essenciais, que destacam-se por ter múltiplas atividades biológicas com potencial farmacológico (Stefanello *et al.,* 2011). Estudos prévios determinaram que esta composição é afetada por diferentes fatores ambientais, como temperatura, disponibilidade hídrica, altitude, nutrientes, entre outros (Burt, 2004).

1.2 *Eugenia uniflora*

O gênero *Eugenia* é membro da subfamília Myrtoideae e contém cerca de 400 espécies no Brasil (Henriques *et al., 1993*). No nível morfológico, as espécies caracterizam-se por serem árvores ou arbustos, apresentarem flores tetrâmeras e pentâmeras, solitárias ou em racemos, cálice aberto ou fechado, ovário bilocular com uma ou duas sementes e embriões com cotilédones (Braga, 1985). A maioria destas espécies são ricas em óleos essenciais e taninos, pelo que são frequentemente utilizadas na medicina popular (Adebajo *et al*., 1989). As espécies deste gênero também se caracterizam pela produção

de frutos comestíveis, como no caso de *E. uniflora*, *E. involucrata*, *E. pyriformis* e *E. neosilvestres*, que são consumidas tanto pelo homem como outros animais (Romagnolo e De Souza, 2006).

E. uniflora, pitanga, nangapiri ou pitangueira é uma espécie nativa da Mata Atlântica e cresce em regiões de clima tropical e subtropical, ocorrendo naturalmente no Brasil (do Rio Grande do Sul ao estado de Pernanbuco), assim como na Argentina, Uruguai e Paraguai (Consolini *et al*., 2002; Bicas *et al*., 2011). Em consequência da sua adaptabilidade aos diferentes tipos de ambientes, foi disseminada amplamente e atualmente pode ser encontrada em diversas partes do mundo como arbusto ou árvore de pequeno porte na restinga, ou como árvore na vegetação ribeirinha (Salgueiro *et al*., 2004). Por este motivo, é considerada como um potencial modelo no aspecto ecológico e genético para a compreensão da adaptabilidade ao meio ambiente e inter-relações entre as plantas e os fatores abióticos.

A pitanga produz frutos com formato semelhante ao de pequenas abóboras de cor laranja ou vermelho obscuro, com aproximadamente 3 cm de diâmetro e oito ranhuras longitudinais, que são consumidos *in natura* ou utilizados na fabricação de sucos, sorvetes e licores (Lim, 2012). A polpa é a principal forma de comercialização, correspondendo aproximadamente a 77% de toda a fruta (Bicas *et al*., 2011). Diferentes estudos prévios reportaram que os frutos tem uma alta concentração de vitamina C, cálcio, fósforo, ferro, vitamina A, riboflavina e niacina (Lorenzi *et al*., 2006). No nível de compostos fitoquímicos, Oliveira *et al*. (2006) identificaram 54 componentes voláteis nos frutos da pitanga. Destes componentes, 29 foram identificados, sendo a maioria monoterpenos (75,3% em massa), incluindo trans-β-ocimene (36,2%), cis-ocimeno (13,4%), β-ocimeno (15,4%) e β pineno (10,3%). No aspecto econômico, a maior região produtora de frutos de pitanga no Brasil é o estado de Pernambuco, com uma produção anual de 1700 toneladas (Bezerra *et al*., 2000).

As sementes de pitanga, que estão geralmente presentes em uma ou duas por fruto, são um recurso de componentes ativos, agentes antibacterianos, compostos fenólicos e fibras dietéticas (Bagetti *et al*., 2009, Luzia *et al*., 2010). Além disso, as sementes de pitanga tem uma alta proporção

de ácidos graxos insaturados (60-70%), sendo 13 a 16% ácidos graxos monoinsaturados e 45-47% de ácidos graxos poliinsaturados (Bageti *et al*., 2009). Devido ao seu efeito na redução da incidência de doenças cardiovasculares, os ácidos graxos poliinsaturados são compostos considerados desejáveis na dieta humana (Lim, 2012).

As folhas de *E. uniflora* são também fontes de compostos ativos e tem sido usadas na medicina popular para o tratamento de diferentes doenças (Consolini *et al*., 1999). Por este motivo, os extratos das folhas da pitanga tem sido foco de vários estudos fitoquímicos, que revelaram a existência de importantes compostos farmacológicos com atividades antioxidantes, antimicrobianas, antihiperglicêmicos, antihipertriglicerimedicos, hipotensores, vasorelaxantes, antivirais, antinociceptivos, hipotérmicos, diuréticos, relacionadas ao sistema nervoso central, anti-inflamatórios, antidiarreicos, contráteis do músculo, tripanocidas, potenciadores de antibióticos e de toxicidade (Lim, 2012).

A maioria dos estudos sobre os compostos voláteis de *E. uniflora* são focados nas folhas (Amorim *et al*., 2009). Os estudos mostram que os óleos essenciais da pitanga diferem na sua composição dependendo da origem geográfica das plantas (Thambi *et al*., 2013). As pitangas obtidas na Nigéria contém uma maior proporção de sesquiterpenos como furanodieno, furanoelemeno, selina-1,3,7(11)-triene-8-one e oxidoselina-1,3,7(11)-triene-8-one (Oguntimein *et al*., 1991), enquanto aquelas obtidas na Argentina tem uma maior proporção de monoterpenos como limoneno, pulogeno, carvono e nerolidol (Lee *et al*., 1997). No caso do Brasil, os principais componentes dos óleos essenciais das pitangas do Rio Grande do Sul são o β-selineno, α-selineno e nerolidol, enquanto as pitangas de Ceará contem selina-1,3,7(11)-triene-8-one e oxidoselina-1,3,7(11)-triene-8-one (Henriques *et al*., 1993). Esta variação observada nestes estudos poderia ser consequência da variação no clima, composição do solo, altitude, temporada de coleta e método usado na extração dos compostos (Costa *et al*., 2009).

No nível genético, Da Costa *et al*. (2008) determinou que a pitanga é uma espécie diploide, com um 2n = 22 cromossomos e com um tamanho de genoma de aproximadamente 245 Mpb. Ao nível molecular poucos estudos

foram realizados por consequência da presença de polifenóis nas folhas das pitangas, que dificultam a extração de DNA (De Almeida *et al*., 2012). Nesse sentido, os trabalhos realizados até o momento são focados no estudo da diversidade genética desta espécie utilizando marcadores RAPDs (Aguiar *et al*., 2013), AFLPs (Margis *et al*., 2002; Salgueiro *et al*., 2004; Nogueira *et al*., 2007; Franzon *et al*., 2010), DNA cloroplástico (Da Cruz *et al*., 2013) e microssatélites (Ferreira-Ramos *et al*., 2008, Ferreira-Ramos *et al*., 2014).

No nível de recursos genômicos, a disponibilidade de sequências de *E. uniflora* é escassa. Até o presente, apenas 36 sequências foram depositadas no banco de dados do National Center for Biotechnology Information (NCBI): nove sequências de marcadores microssatélites (Ferreira-Ramos *et al*., 2008) e 27 sequências de genes obtidos de estudos filogenéticos (Biffin *et al*., 2006; Clausing e Renner, 2001; Da Cruz *et al*., 2013; Kitson *et al*., 2013; Luca *et al*., 2007; Rutschmann *et al*., 2007; Soh e Parnell, 2011; Van der Merwe *et al*., 2005; Wilson *et al*., 2001). O número de sequências disponíveis da pitanga é muito pequeno em comparação do número de sequências disponíveis de *Eucalyptus grandis*, espécie com genoma sequenciado recentemente (Myburg *et al*., 2014).

1.3 Sínteses de terpenóides em plantas

Os óleos essenciais são compostos que, por suas características químicas, podem ser classificados como terpenóides e não terpenóides. Estes compostos são substâncias voláteis, obtidas mediante processos físicos a partir de espécies vegetais aromáticas. Correspondem a misturas complexas de mais de 100 componentes que geralmente são compostos alifáticos de baixo peso molecular, monoterpenos, sesquiterpenos e fenilpropanos (Villar del Fresno, 1999). Estas substâncias aromáticas voláteis são sintetizadas em diferentes órgãos das plantas e armazenadas nos canais secretores das mesmas (Hoffmann *et al*., 2003). Os óleos essenciais são comumente encontrados em espécies das famílias Apiaceae, Asteraceae, Lamiaceae, Lauraceae, Umbelíferaceae e Myrtaceae (Bruneton, 2001).

Os terpenóides, também conhecidos como terpenos ou isoprenóides, constituem uma grande classe de produtos naturais vegetais com uma alta diversidade funcional e mais de 20.000 membros (Schwab, 2003). Os terpenóides servem como precursores da biossíntese de metabólitos essenciais para as plantas, incluindo aqueles envolvidos na regulação do crescimento (giberelinas, ácido abscísico, e strigolactonas), estabilização da membrana (esteróis), fotossíntese (carotenóides e cadeia lateral fitol da clorofila) e coenzimas do transporte de elétrons (ubiquinona e plastoquinona) (Croteau *et al*., 2000). Todos os terpenóides são baseados em unidades de 5 carbonos (C5), como o isopentenil pirofosfato (IPP) ou difosfato de dimetilalilo (DMAPP), e seus esqueletos carbonados são construídos a partir da união de dois ou mais destas unidades (Liu *et al*., 2014). Os terpenóides estão divididos nos voláteis monoterpenos (C10) e sesquiterpenos (C15), nos menos voláteis diterpenos (C20), e nos não voláteis triterpenos e esteróis (C30), pigmentos carotenoides (C40) e poliisopropenos (Cn) (Harborne, 1984). A maioria dos terpenóides naturais tem estruturas cíclicas com um ou mais grupos funcionais, por consequência de que as últimas etapas de síntese envolvem ciclização, oxidação ou outra modificação estrutural (Villar del Fresno, 1999).

Em plantas, o IPP e DMAPP são sintetizados por meio de duas vias compartimentalizadas: a via do mevalonato e a via do metileritritolo fosfato (MEP). A via do mevalonato opera no retículo endoplasmático e peroxisomas (Lange e Ahkami, 2013). A condensação de 3 unidades de acetil CoA conduz à síntese de 3-hidroxi-3-metilglutaril CoA, que produz ácido mevalônico em outra etapa mais adiante. O ácido mevalônico é convertido em isopentenilo difosfato através de um processo de fosforilação e descarboxilação. O 3-hidroxi-3-metilglutaril CoA redutase catalisa a redução da 3-hidroxi-3-metilglutaril CoA a mevalonato ácido (Rodwell *et al*., 2000). Em *Arabidopsis thaliana*, o mevalonato-5-difosfato é produzido pela fosforilação do ácido mevalônico, realizada pelas enzimas mevalonato cinase e fosfomevalonato cinase (Lluch *et al*., 2000). Mais tarde, a mevalonato-5-difosfato descarboxilase catalisa a conversão do mevalonato-5-difosfato em isopentenilo difosfato, o qual é o produto final da via do mevalonato (Dhe-Paganon *et al*., 1994).

A via do metileritritolo fosfato começa nos plastídios através da condensação do ácido pirúvico e gliceraldeído-3-fosfato, que conduz à síntese de 1-desoxi-D-xilulose 5-fosfato. Esta reação é catalisada pela enzima 1-desoxi-D-xilulose 5-fosfato sintase (Rodríguez-Concepción e Boronat, 2002). Posteriormente, o 1-desoxi-D-xilulose 5-fosfato é reduzido em 2-C-metil-D-eritritol 4-fosfato pela 1-desoxi-D-xilulose redutoisomerase (Takahashi *et al*., 1998). A conjugação de 2-C-metil-D-eritritol 4-fosfato e 4-citidina 5-fosfato leva à formação da 4-citidina 5-fosfo-2-C-metil eritritol e a reação é catalisada pela enzima 2-C-metil-D-eritritol 4-fosfato citidililtransferase. O 4-citidina 5-fosfo-2-C-metil eritritol é convertida em 2-C-metil eritritol 2,4-ciclodifosfato pela enzima 2-C-metil eritritol 2,4-ciclodifosfato sintase (Calisto *et al*., 2007). Todas as enzimas da via 2C-metil-D-eritritol-4-fosfato estão localizadas nos plastídios (Hsieh *et al*., 2008). Na via do 1-desoxi-D-xilulose 5-fosfato, a síntese do hidroximetilbutanal 4-difosfato é realizada a partir do 2-C-metil eritritol 2,4-ciclodifosfato e a reação é catalisada pela hidroximetilbutanal 4-difosfato sintase. Posteriormente, o hidroximetilbutanal 4-difosfato é diretamente convertido em IPP e DMAPP pelas enzimas isopentinilo difosfatase e difosfato de dimetilalilo sintase, respetivamente (Cunningham *et al*., 2000).

A condensação do IPP e DMAPP é catalisada pelas preniltransferases (PTs) e tem como resultado três principais intermediários da via dos terpenóides: geranil pirofosfato (GPP), farnesil difosfato (FDP) e geranilgeranil pirofosfato (GGPP) (Koyama e Ogura, 1999). As centenas de esqueletos básicos típicos dos terpenóides das plantas são formados a partir de DMAPP, GPP, FDP ou GGPP pelas terpeno sintases (TPS) (Davis e Croteau, 2000). As TPS formam o hemiterpeno isopreno (C5) a partir do DMAPP (Miller *et al*., 2001), monoterpenos (C10) a partir do GPP (Wise e Croteau, 1999), sesquiterpenos (C15) a partir do FDP (Cane, 1999), e diterpenos (C20) a partir de GGPP (MacMillan e Beale, 1999). Estas sintases, com a exceção de copalyl difosfato (CPP) sintase, funcionam através da geração de intermediários de carbocátion a partir dos respectivos substratos de 5C por uma reação dependente de íon metálico divalente (Davis e Croteau, 2000).

Níveis comparativos de identidade entre terpeno sintases de plantas já conhecidas levou à designação de subgrupos de terpeno sintases

diferenciados por um mínimo de 40 % de identidade dos aminoácidos entre os membros e valores menores entre os subgrupos (Bohlmann *et al*., 1998a). Inicialmente as terpeno sintases foram divididas em seis subgrupos designados de TPS-a até TPS-f. O subgrupo TPS-a compreende principalmente sesquiterpenos sintases das angiospermas enquanto o subgrupo TPS-b é dominado pelas monoterpeno sintases das angiospermas. Os subgrupos TPS-c e TPS-e agrupam diterpenos sintases envolvidas na biossíntese de giberelina, e incluem a (-)-copalyl difosfato sintase (Ait-Ali *et al*., 1997) e kaurene sintase (Yamaguchi *et al*., 1996). O subgrupo TPS-d está inteiramente composto por terpeno sintases de gimnospermas que produzem principalmente monoterpenos. O subgrupo TPS-f compreende as monoterpeno sintases de angiospermas que possuem um domínio conservado N-terminal de 200 aminoácidos de função desconhecida (Bohlmann *et al*., 1998b). Mais recentemente, um trabalho feito em flores de *Antirrhinum majus* levou à descoberta de duas mirceno sintases e uma (E)-β-ocimeno sintase que, junto com a enzima AtTPS14 de *Arabidopsis,* formam o novo subgrupo TPS-g (Dudareva *et al*., 2003).

A ampla gama de atividades das TPS em plantas foram reveladas pela primeira vez em estudos com extratos vegetais brutos e enzimas nativas purificadas (Wise e Croteau, 1999). Na atualidade, o isolamento dos genes das TPS e sua expressão heteróloga fornecem a melhor evidência para a caracterização da função e processo catalítico destas enzimas. Como resultado deste tipo de estudos, observou-se que as TPS podem formar produtos multiméricos a partir do mesmo substrato (Degenhardt *et al*., 2009). Recentemente, estudos no nível genômico permitiram a identificação dos genes das TPS em plantas com genomas sequenciados. Por exemplo, 32, 69 e 120 potenciais genes de TPS foram relatados em *A. thaliana*, *Vitis vinifera* e *E. grandis*, respetivamente (Aubourg *et al*., 2002; Grattapaglia *et al*., 2012).

No caso da síntese dos triterpenos, a cadeia linear do esqualeno é derivada do acoplamento redutivo de duas moléculas de farnesil pirofosfato (FPP) pela esqualeno sintase (Singh e Sharma, 2014). O esqualeno posteriormente é oxidado pela enzima esqualeno epoxidase para gerar o 2,3-oxidoesqualeno. Finalmente, o 2,3-oxidoesqualeno é convertido em triterpenos

pelos membros da família das triterpenos sintases ou oxidoesqualeno ciclases (OSC) (Phillips *et al*., 2006; Jenner *et al*., 2005). Diferentes tipos de oxidoesqualeno sintases foram isolados em várias espécies de plantas incluindo: lanosterol sintase (Baker *et al*., 1995), cicloartenol sintase (Kawano *et al*., 2002), lupeol sintase (Hayashi *et al*., 2004) e b-amirina sintase (Iturbe-Ormaetxe *et al*., 2003). Além destas sintases, algumas outras triterpenos sintases foram também caracterizadas em outras diferentes espécies de plantas (Basyuni *et al*., 2006; Shibuya *et al*., 2007).

Ao nível genômico, 13 genes da família OSC foram descritos em *A. thaliana* (Singh e Sharma, 2014). O primeiro gene OSC de *A. thaliana* a ser clonado foi a cicloartenol sintase (CAS1), usando uma estratégia de complementação funcional (Corey *et al*., 1993). Posteriormente foi identificado outro gene que codifica uma lanosterol sintase, que está envolvida na biossíntese de fitoesteróis (Ohyama *et al*., 2009). Os outros 11 membros da família OSC em *A. thaliana* produzem uma ampla diversidade de esqueletos de triterpenos (mais de 40 no total), confirmando a notável diversidade química derivada a partir de um único substrato (Phillips *et al*., 2006; Abe, 2007; Morlacchi *et al*., 2009). Um estudo prévio realizado no genoma de *Oryza sativa* L. ssp. Japonica cv Nipponbare identificou 12 genes OSC (Inagaki *et al*., 2011). Além disso, a anotação automatizada dos genomas de *Sorghum bicolor* e *Brachypodium distachyon* reporta a presença de diferentes genes OSC de função desconhecida (Paterson *et al*., 2009). No entanto, os genomas de plantas inferiores, como *Chlamydomonas reinhardtii* e *Physcomitrella patens,* apresentam apenas um gene, provavelmente necessário para a biossíntese de esteróis (Merchant *et al*., 2007; Desmond e Gribaldo, 2009).

1.4 Os miRNAs e sua importância na regulação gênica

Os miRNAs são uma classe de pequenos RNAs de 19 a 24 nucleotídeos que desempenham uma importante função na regulação postranscricional da expressão gênica por ligação ao mRNA alvo com alta complementariedade (Bartel, 2004). O impacto mais importante das atividades dos miRNAs incluem: embriogênese, desenvolvimento floral, morfogênese da folha, padronização,

transição de fase, desenvolvimento das anteras e reprodução (Ramesh *et al*., 2014). Os miRNAs também estão relacionados com as respostas das plantas aos estresses bióticos e abióticos (Sunkar e Zhu, 2004; Navarro *et al*., 2006; Sunkar *et al*., 2012). Inicialmente os miRNAs eram identificados usando métodos de clonagem direta (Wang *et al.,* 2007). Nos últimos anos, o número de miRNAs identificados em plantas foi aumentado pelo uso do sequenciamento de nova geração (NGS) e análises bioinformáticas (Liang *et al*., 2010; Xin *et al*., 2010).

A biogênese dos miRNAs ocorre dentro do núcleo, em regiões chamadas de corpos D (Fang e Spector, 2007). O processo é iniciado a partir RNAs de fita simples longas, chamadas de transcritos de miRNA primário (pri-miRNAs). Os pri-miRNAs são caracterizados como estruturas em grampo (stem-loop) imperfeitas, que são geradas a partir da atividade da RNA polimerase II (Lee *et al*., 2004). A conversão dos pri-miRNAs, através de um precursor de miRNAs (pre-miRNAs), para um duplex miRNA:miRNA* é coordenada pela atividade de diferentes famílias de proteínas, incluindo a RNAse III e Dicer like 1 (DCL-1) (Xie *et al*., 2005). Os pre-miRNAs de plantas, os quais são relativamente longos comparados aos pre-miRNAs de origem animal, são cortadas no pequeno RNA maduro de fita dupla (miRNA:miRNA*) e transportados ao citoplasma pela EXPORTIN-5 (Kurihara e Watanabe, 2004). Neste processo, a enzima DCL-1 atua em conjunto com as proteínas de união ao RNA de fita dupla, como o HYL1 e SERRATE dentro do núcleo para produzir miRNAs maduros (Han *et al*., 2004). O miRNA é metilado e poliuridinilado pela enzima HEN I que permite uma proteção contra degradação por exonucleases (Kurihara *et al*., 2006). No citoplasma, os miRNAs maduros de 21 nt de comprimento são recrutados pelo complexo de silenciamento induzido pelo RNA (RISC), promovendo a clivagem da sequência de mRNAs alvos (Bartel, 2004) ou a repressão da tradução (Brodersen *et al*., 2008).

Atualmente, um total de 8.496 miRNAs maduros já foram descritos para 73 espécies de Viridiplantae e depositados no banco de dados miRBase em sua versão 21.0 (http://www.mirbase.org/). Os miRNAs maduros de plantas são designados com as letras iniciais miR, seguidos por um número (Reinhart *et al*., 2002). O banco de dados do miRBase usa abreviadamente três letras para

designar as espécies, como por exemplo Ath-miR172 no caso do *A. thaliana* e Osa-miR160 no caso de *O. sativa*. Seus genes correspondentes estão em maiúsculas e itálico: miR156a refere-se ao miRNA maduro e *MIR156a* refere-se ao gene que codifica esse miRNA (Reinhart *et al*., 2002; Griffiths-Jones, 2004). No caso da existência de mais de um locus para o mesmo miRNA, os nomes são numericamente sufixados com letras, como no caso do miR156a, miR156b e miR156c (Griffiths-Jones *et al*., 2006).

Diferentes estudos mostram que muitos miRNAs são altamente conservados ao longo da evolução e podem ser encontrados de musgos até plantas superiores, enquanto um número significativo de miRNAs não conservados também têm sido identificados como espécies-específicos (Jones-Rhoades *et al*., 2006). No caso dos miRNAs não conservados, a maioria deles são expressos especificamente em um determinado tecido ou em um estádio específico de desenvolvimento (Allen *et al*., 2004). Igualmente, um estudo desenvolvido por Cuperus *et al*. (2011) sugere que os miRNAs não conservados tem um passado evolutivo recente e foram integrados em redes regulatórias específicas de plantas. Consequência da sua influência vantajosa na regulação de funções e vias específicas de tecidos.

1.5 Obtenção de transcritomas *de novo* através de RNA-seq

O RNA-seq é um método de sequenciamento de alto rendimento desenvolvido recentemente, que produz milhões de sequências curtas de RNA que são posteriormente montados em transcritos, usando ou não um genoma de referência (Kumar *et al*., 2014). Em plantas, esta metodologia acelerou a compreensão dos complexos padrões de transcrição e propiciou a quantificação da expressão gênica em diferentes tecidos ou estágios de desenvolvimento (Zenoni *et al*., 2010). Além disso, o método do RNA-seq permite a identificação de novas regiões transcritas e isoformas resultantes do *splicing* e de polimorfismos de nucleotídeo único (SNP) (Cloonan *et al*., 2008; Wilhelm *et al*., 2010). Além disso, na ausência de um genoma de referência, o sequenciamento do transcritomas não é apenas usado na identificação de transcritos envolvidos em um processo biológico específico, mas pode ser

usado também como uma alternativa na obtenção de um número grande de marcadores moleculares como microssatélites a partir das sequências não redundantes dos transcritos (Parida *et al*., 2006).

Nas duas décadas passadas, o método automatizado de Sanger foi denominado como a tecnologia de primeira geração e as novas tecnologias da atualidade são consideradas como as tecnologias de próxima geração (NGS) (Metzker, 2009). As NGS mais amplamente utilizadas são o Roche 454 Life Sciences, Applied Biosystems SOLiD e Illumina Genome Analyzer, sendo esta última a tecnologia mais amplamente usada pela alta cobertura de leituras geradas por corrida. Por causa da geração de milhões de leituras, as ferramentas bioinformáticas disponíveis na atualidade para realizar uma montagem *de novo* precisam presentam duas dificuldades técnicas (Clarke *et al*., 2013). Primeiro, precisa-se de altos requerimentos em velocidades computacionais e de alta capacidade de memória, uma vez que os alinhamentos emparelhados feitos entre as leituras requer o uso de muitos gigabytes para os dados de entrada e intermediários. Segundo, os genes de células eucarióticas codificam diferentes transcritos que compartilham exons entre elas, o que resulta em alguns fragmentos de RNA com uma incorreta concatenação de leituras.

Atualmente existem dois tipos principais de algoritmos de montagem para leituras obtidas com NGS: gráficos de sobreposição e gráficos de Bruijn (Li *et al*., 2012). O gráfico de sobreposição está baseado no alinhamento pareado das leituras (Myers, 1995). Neste tipo de gráfico, cada *node* representa uma leitura e a borda entre dois *nodes* indica que duas leituras tem sequências que sobrepõem-se. Depois de algumas etapas de simplificação do gráfico, que consistem na remoção dos *nodes* e bordas transitivas, uma cadeia de *nodes* representa a sequência de um *contig* ou transcrito. No entanto, a etapa de alinhamentos pareados de leituras obtidas por NGS fazem o uso deste tipo de gráficos, que é muito exigente ao nível computacional. As ferramentas de bioinformática baseadas no gráfico de sobreposição mais usadas são Mira (Chevreux *et al*., 2004) e Newbler (Margulies *et al*., 2005).

Pela necessidade de contar com um algoritmo gráfico muito mais rápido, foi desenvolvido o gráfico de Bruijn (Idury e Waterman, 1995). Neste tipo de

gráfico, as leituras são quebradas em subsequências de tamanho k, chamadas k-mer, as quais são usadas para a construção do gráfico de forma linear. As três ferramentas de bioinformática baseadas neste tipo de gráfico são ABySS (Simpson *et al*., 2009), Trinity (Grabherr *et al*., 2011) e Velvet (Zerbino e Birney, 2008). A ferramenta Oases (Schulz *et al*., 2012), desenvolvida especialmente para a montagem de transcritomas, é como uma extensão do Velvet e trabalha considerando o *splicing* alternativo. A principal desvantagem dos gráficos de Bruijn é a perda da informação das leituras quando são quebradas em k-mer (Clarke *et al*., 2013). Desta maneira, quando dois genes têm uma sequência compartilhada de um comprimento maior que k, o gráfico de Bruijn pode conectar incorretamente as leituras destes dois genes.

A montagem de um transcritoma pode gerar muitas sequências de transcritos não identificados. Por esse motivo, com a finalidade de interpretar biologicamente estes dados, os transcritos necessitam ser associados com termos e nomes que fornecem informações sobre suas funções. A anotação destas sequências pode incluir busca por similaridade contra outros genes e proteínas com funções conhecidas. Esta busca de similaridade é feita dentro de bancos de dados de sequências como Genbank do NCBI (Benson *et al*., 2012) ou UniProt (Consortium UniProt, 2013), e são escolhidas como referências as sequências com a mais alta similaridade. Alternativamente, os transcritos podem ser anotados funcionalmente através da procura nas suas sequências de domínios e motivos conservados dentro de bases de dados, como do InterPro do EBI (Hunter *et al*., 2011). Outras alternativas de anotação funcional incluem o Gene Ontology (GO) (Ashburner *et al*., 2000), números da Comissão de enzimas (EC) (Schomburg *et al*., 2004) e vias da KEGG (Kanehisa *et al*., 2014).

No caso da família Myrtaceae, os únicos estudos de sequenciamento de transcritomas usando tecnologias de sequenciamento de nova geração foram feitos em espécies do gênero Eucalyptus (Grattapaglia *et al*., 2012).

Capítulo 2

Objetivos

2. *Objetivos*

2.1 Objetivo geral

Gerar recursos genômicos de *Eugenia uniflora* usando tecnologias de sequenciamento de nova geração para a identificação de genes e miRNAs envolvidos em diferentes processos fisiológicos e metabólicos característicos da pitanga. Os dados obtidos neste estudo deverão servir como uma referência de base e de alta qualidade para futuros estudos de genômica funcional na pitanga e outras diversas espécies de interesse da família Myrtaceae.

2.2 Objetivos específicos

1. Construir e sequenciar bibliotecas de mRNA e pequenos RNAs a partir de folhas de *E. uniflora*.

2. Identificar e verificar os miRNAs maduros e precursores de miRNAs expressos nas folhas de *E. uniflora*.

3. Montar *de novo* e anotar o transcriptoma de *E. uniflora*.

4. Identificar e verificar a expressão das terpeno sintases e oxidoesqualeno ciclases expressas nas folhas de *E. uniflora*.

Capítulo 3

Identification of microRNAs from *Eugenia uniflora* by high-throughput sequencing and bioinformatics analysis

Frank Guzman[1,2], Mauricio P. Almerão[2], Ana P. Korbes[1], Guilherme Loss-Morais[2], Rogerio Margis[1,2,3]

1 PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil,

2 PPGBCM, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil,

3 Departamento de Biofísica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil

Artigo publicado na PLoS ONE (2012)

OPEN ACCESS Freely available online

PLOS ONE

Identification of MicroRNAs from *Eugenia uniflora* by High-Throughput Sequencing and Bioinformatics Analysis

Frank Guzman[1,2], Mauricio P. Almerão[2], Ana P. Körbes[1], Guilherme Loss-Morais[2], Rogerio Margis[1,2,3]*

1 PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **2** PPGBCM, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **3** Departamento de Biofisica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil

Abstract

Background: microRNAs or miRNAs are small non-coding regulatory RNAs that play important functions in the regulation of gene expression at the post-transcriptional level by targeting mRNAs for degradation or inhibiting protein translation. *Eugenia uniflora* is a plant native to tropical America with pharmacological and ecological importance, and there have been no previous studies concerning its gene expression and regulation. To date, no miRNAs have been reported in Myrtaceae species.

Results: Small RNA and RNA-seq libraries were constructed to identify miRNAs and pre-miRNAs in *Eugenia uniflora*. Solexa technology was used to perform high throughput sequencing of the library, and the data obtained were analyzed using bioinformatics tools. From 14,489,131 small RNA clean reads, we obtained 1,852,722 mature miRNA sequences representing 45 conserved families that have been identified in other plant species. Further analysis using contigs assembled from RNA-seq allowed the prediction of secondary structures of 25 known and 17 novel pre-miRNAs. The expression of twenty-seven identified miRNAs was also validated using RT-PCR assays. Potential targets were predicted for the most abundant mature miRNAs in the identified pre-miRNAs based on sequence homology.

Conclusions: This study is the first large scale identification of miRNAs and their potential targets from a species of the Myrtaceae family without genomic sequence resources. Our study provides more information about the evolutionary conservation of the regulatory network of miRNAs in plants and highlights species-specific miRNAs.

Citation: Guzman F, Almerão MP, Körbes AP, Loss-Morais G, Margis R (2012) Identification of MicroRNAs from *Eugenia uniflora* by High-Throughput Sequencing and Bioinformatics Analysis. PLoS ONE 7(11): e49811. doi:10.1371/journal.pone.0049811

Editor: Abidur Rahman, Iwate University, Japan

Received June 18, 2012; **Accepted** October 17, 2012; **Published** November 15, 2012

Funding: This research was supported by grants from FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico). FG received a PhD fellowship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior). APK, MPA and RM have PNPD/CAPES, PDJ/CNPq Research/CNPq fellowships, respectively. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rogerio.margis@ufrgs.br

Introduction

Eugenia uniflora is a tropical fruit tree native to South America. The shrubby tree produces edible cherry-like fruits, which are locally known as pitanga or the Brazilian cherry. This species belongs to the Myrtaceae family, which is characterized by the presence of tannins, flavonoids, monoterpenes and sesquiterpenes whose presence and concentration varies between specimens from different geographical locations [1–3]. Extracts from pitanga leaves contain interesting biological properties that have been reported in several studies, and pitanga juice is used in folk medicine as a diuretic, antirheumatic, antipyretic, antidiarrhetic and antidiabetic [4–9]. *E. uniflora* is also an important ecological model to study because it grows in areas of medium and large levels of rainfall and can also be found in different vegetation types and ecosystems [10]. The variation in the metabolite concentration and the adaptability to different environments observed in *E. uniflora* indicating that these are the result of the transcriptional regulation of many genes involved in metabolic and signaling pathways.

MicroRNAs (miRNAs) are small non-coding regulatory RNAs widely found in unicellular and multicellular organisms that act as regulators of gene expression at the post-transcriptional level on genes containing miRNA target sites [11]. Mature miRNAs are single-stranded RNA molecules of approximately 21 nucleotides (nt) in length processed from a precursor molecule (pre-miRNA) [12]. To regulate protein-coding genes, the mature miRNA binds with perfect or imperfect complementarity to sites in the 5′ or 3′ untranslated regions (UTR) or coding sequences (CDS) of genes, which leads to mRNA degradation or translation inhibition [13–14]. In plants, miRNAs have diverse biological functions and are involved in the regulation of optimal growth and development as well as other physiological processes, including abiotic and biotic stress responses [15]. Several studies showed that many miRNAs are conserved across different plant families [16–17]. However, family- and species-specific miRNAs that are expressed in lower

levels and probably have evolved more recently have been reported [18].

In the present study, in order to evaluate the importance of miRNAs in the regulation of gene expression and metabolic pathways in *E. uniflora*, we constructed small RNA (sRNA) and polyA RNA-seq libraries from leaves and sequenced the libraries with high throughput Solexa technology. The sequencing data were analyzed to identify conserved and novel miRNAs and their respective targets. This work represents the first report of miRNAs identified in Myrtaceae.

Methods

Plant Material and RNA Isolation

Total RNA was isolated from *E. uniflora* leaves using the CTAB method [19]. RNA quality was evaluated by electrophoresis on a 1% agarose gel, and quantification was determined using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

Deep Sequencing

Total RNA (>10 μg) was sent to Fasteris SA (Plan-les-Ouates, Switzerland) for processing. One sRNA library was constructed and sequenced using the Illumina HiSeq2000 platform. Briefly, the construction of the sRNA library consisted of the following successive steps: acrylamide gel purification of the RNA bands corresponding to a size range of 20–30 nt; ligation of the 3p and 5p adapters to the RNA in two separate subsequent steps, each followed by acrylamide gel purification; cDNA synthesis followed by acrylamide gel purification; and a final step of PCR amplification to generate a cDNA colony template library for Illumina sequencing.

The polyadenylated transcript sequencing (RNA-seq) was performed using the following successive steps: poly-A purification; cDNA synthesis using a poly-T primer shotgun method to generate inserts of 500 nt, 3p and 5p adapter ligations; pre-amplification; colony generation; and sequencing. The Illumina output data included sequence tags of 100 bases.

Accession Numbers

Sequencing data are available at the NCBI Gene Expression Omnibus (GEO) ([http://www.ncbi.nlm.nih.gov/geo]). The accession number GSE38212 contains the sequence data from the RNA-seq and sRNA libraries derived from *E. uniflora* leaves.

Data Analysis

The overall procedure for analyzing Illumina small libraries is shown in Figure S1. All low quality reads with FASTq values below 13 were removed, and 5′ and 3′ adapter sequences were trimmed using the Genome Analyzer Pipeline (Illumina) at Fasteris SA. The remaining low quality reads with 'n' were removed with PrinSeq script [20]. Sequences shorter than 18 nt and larger than 25 nt were excluded from further analysis. Small RNAs derived from Viridiplantae rRNAs, tRNAs, snRNAs and snoRNAs deposited at the tRNAdb [21], SILVA rRNA [22], and NONCODE v3.0 [23] databases and from Rosales mtDNA and cpDNA sequences deposited at the NCBI GenBank database [(http://ftp.ncbi.nlm.nih.gov)] were identified by mapping with Bowtie [24].

After cleaning the data (low quality reads, adapter sequences), the RNA-seq data were assembled into contigs using the CLC Genome Workbench version 4.0.2 (CLCbio, Aarhus, Denmark) algorithm for *de novo* sequence assembly using the default parameters (similarity = 0.8, length fraction = 0.5, insertion/deletion cost = 3, mismatch cost = 3). In total, 170,568 contigs were assembled and used as a reference for the discovery of pre-miRNA and target sequences.

Identification of Conserved and Novel miRNAs

In order to determine conserved plant miRNAs, small RNA sequences were aligned with known non-redundant Magnoliophyta miRNAs deposited at miRBase (Release 18, November 2011) using Bowtie. Complete alignment of the sequences was required, and no mismatches were allowed. To search for novel miRNAs, small RNA sequences were matched against contigs obtained through *de novo* assembly of transcripts from mRNA sequences of *E. uniflora* leaves using SOAP2 [25]. The SOAP2 output was filtered with an in-house filter tool (FilterPrecursor) to identify candidate sequences as miRNA precursors using an anchoring pattern of one or two blocks of aligned small RNAs with perfect matches. The selected candidate precursors were manually inspected using the software Tablet [26] to visualize the presence of the anchoring pattern. As miRNA precursors have a characteristic hairpin structure, the next step to select candidate sequences included secondary structure analysis by RNAfold with the annotation algorithm from the UEA sRNA toolkit [27]. In addition, perfect stem-loop structures should have the miRNA sequence at one arm of the stem and a respective anti-sense sequence at the opposite arm. In the present study miRNAs were named in two different ways: (i) miR000, when corresponding to a family with two or more miRNA loci and (ii) MIR000, to design a single locus. Finally, precursor candidate sequences were checked using the BLAST algorithm from miRBase (www.mirbase.org).

Validation of miRNA by RT-PCR

In order to validate predicted miRNAs, a series of RT-PCR were performed in RNA isolated from leaves of three individuals of *E. uniflora* occurring in the Grumari native protected area in Rio de Janeiro, Brazil. Among the analyzed miRNAs seventeen corresponded to conserved miRNAs (eun-MIR156, eun-MIR159, eun-MIR160, eun-MIR166, eun-MIR167-1, eun-MIR167-2, eun-MIR167-3, eun-MIR167-4, eun-MIR395, eun-MIR396-1, eun-MIR396-2, eun-MIR397-1, eun-MIR397-2, eun-MIR482-1, eun-MIR482-2, eun-MIR530, eun-MIR827) and ten were novel miRNAs (eun-MIR001-1, eun-MIR001-2, eun-MIR004-2, eun-MIR005, eun-MIR006, eun-MIR008, eun-MIR009, eun-MIR012, eun-MIR013, eun-MIR014). The stem-loop primer, used for miRNA cDNA synthesis, was designed according to Cheng *et al.* [28]. The forward miRNAs primers were designed based on the full mature miRNA sequences and the reverse primer was the universal reverse primer for miRNA. The RT-PCR was performed according the conditions used by Kulcheski *et al.* [29]. Briefly, reactions were completed in a volume of 20 μL containing 10 μL of diluted cDNA (1:100), 0.025 mM dNTP, 1X PCR Buffer, 3 mM MgCl2, 0.25 U Hot Start Taq DNA Polymerase (Promega) and 200 nM of each reverse and forward primer. Samples were analyzed in biological triplicate in a 96-well plate, and a no-template control was included. The PCR conditions were performed in an ABI 7500 Real-Time PCR System (Applied Biosystems). The PCR products were resolved on a 2% agarose gel and analyzed using Quantity One software (Bio-Rad).

Prediction of miRNA Targets

Previously assembled mRNA contigs were clustered using the Gene Indices Clustering Tools (http://compbio.dfci.harvard.edu/tgi/software/) [30] to reduce any sequence redundancy. The clustering output was passed to the CAP3 assembler [31] for

multiple alignment and consensus building. Contigs that did not reach the set threshold and fell into any assembly remained as a list of singletons.

The prediction of target genes for the most abundant mature miRNAs from the conserved and novel pre-miRNAs was performed by psRNAtarget [32]. The program uses a 0–5 scale to indicate the complementarity between miRNA and their target, with the smaller numbers representing higher complementary and zero corresponding to a perfect complementation. Default parameters with an expectation value of 4 and *E. uniflora* assembled unigenes longer than 600 bp were used. Candidate RNA sequences were then annotated by assignment of putative gene descriptions based on sequence similarity with previously identified genes annotated with details deposited in the protein database of NR and the Swiss-Prot/Uniprot protein database using BLASTx implemented in blast2GO v2.3.5 software [33]. The annotation was improved by the analysis of conserved domains/families using the InterProScan tool and Gene Ontology terms as determined by the GOslim tool from blast2GO software. At the same time, the orientations of the transcripts were obtained from BLAST annotations.

Finally, to verify if the genes targeted by the identified miRNAs regulate any metabolic pathways involved in the secondary metabolites synthesis, we obtained the enzyme EC numbers for each target gene from the blast2GO annotation. These codes were uploaded to iPATH2 server [34] to generate metabolic pathway maps.

Results

E. uniflora RNA Library Sequencing

To identify conserved and novel miRNAs in *E. uniflora*, sRNA library was constructed from leaves and sequenced using Solexa high-throughput technology. After removing low quality sequences, those without inserts, or those with adapter contaminants or lengths outside of the 18–25 nt range, a total of 14,849,131 reads were obtained (Table 1). The number of reads with different lengths in the redundant and non-redundant sRNA datasets is shown in Figure 1 and Table S1. The most abundant sRNA species contained 21 nt, whereas the highest sequence diversity was observed in the 24-nt fraction. Approximately 6.55% of the reads matched other types of non-coding sRNAs, such as rRNAs, tRNAs, snRNAs or snoRNAs, and 9.45% matched organellar DNA (Table 2).

As there is no genome sequence available for *E. uniflora*, we sequenced the mRNA transcriptome of the *E. uniflora* leaf for use as a reference sequence in further analyses. The pooled mRNA-seq yielded 16,759,528 reads, which were imported into the CLC Genomics Workbench and *de novo* assembled into 170,568 contigs with an average length of 306 bp. Contigs and non-assembled

Table 1. Summary of data from sequencing of *E. uniflora* small RNA library.

Type	Number of reads	Percentage (%)
Total reads*	14,849,131	100
18–25 nt	12,759,506	86
<18 nt	1,554,975	10
>25 nt	534,650	4

*Reads with high quality with lenghts of 1 to 44 nt.
doi:10.1371/journal.pone.0049811.t001

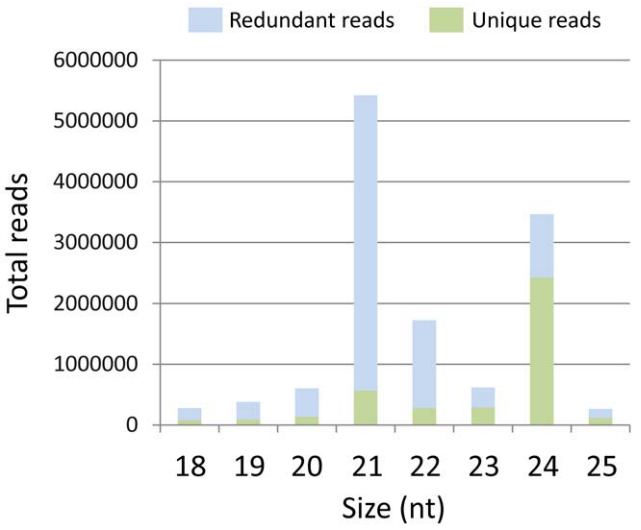

Figure 1. Length distribution and diversity of small RNA reads in the *E. uniflora* leaf library.
doi:10.1371/journal.pone.0049811.g001

reads with minimum lengths of 100 bp were further considered. The contigs ranged in size between the minimum set threshold of 100 bp and 7,808 bp (N50 = 447 bp), with 22,308 contigs more than 500 bp in length.

Identification of Conserved miRNAs in *E. uniflora*

There are 4,677 miRNAs from 47 Magnoliophyta species deposited in miRBase. To identify conserved miRNAs in *E. uniflora*, the small RNA library was matched against a set of 2,585 unique, mature plant miRNA sequences from the database. In total, 1,852,722 reads perfectly matched 204 known miRNAs (Figure 2 and Table S2). All identified sequences are distributed in 45 miRNA families, with an average of approximately 4 miRNA members per family. The largest family was miR166 with 21 members, which include isoforms found in several plant species. The miR156 (19 members), miR396 (15 members) and miR395 (14 members) families were the second, third and fourth largest miRNA families, respectively. Of the remaining miRNA families, 23 contained 2 to 10 members, and 18 were represented by a single member (Figure 2).

Table 2. Categorization of *E. uniflora* noncoding and organellar small RNAs*.

Small RNA type	Number of reads	Percentage (%)
miRNA	1,852,722	14.52%
rRNA	765,989	6.00%
tRNA	67,491	0.53%
snRNA	1,555	0.01%
snoRNA	859	0.01%
mtRNA	159,106	1.25%
cpRNA	1,046,305	8.20%
Other sRNA	8,865,479	69.48%

*18–25 nt reads considered.
doi:10.1371/journal.pone.0049811.t002

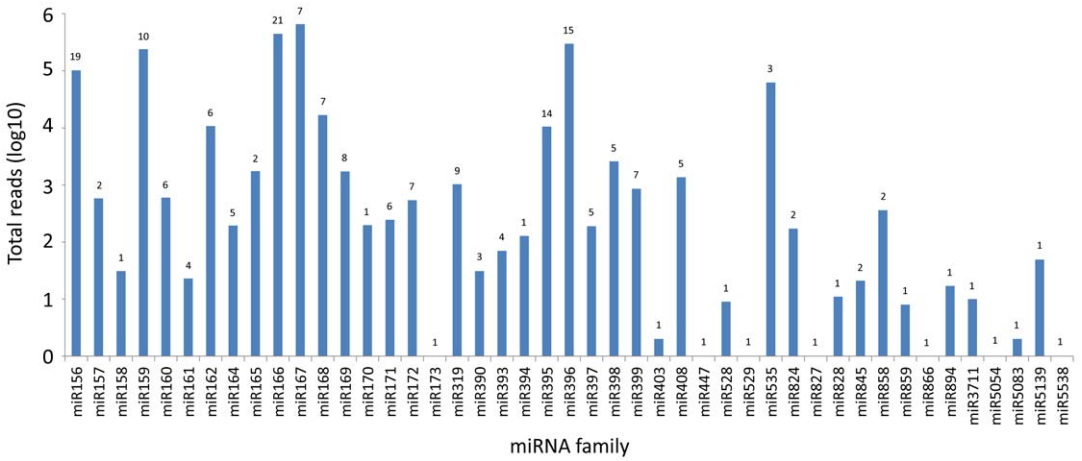

Figure 2. Number of identified miRNAs in each conserved miRNA family in plants. The values above the bars indicate the number of members identified in each conserved miRNA family.
doi:10.1371/journal.pone.0049811.g002

With respect to the abundance of each miRNA family, the frequencies varied from 1 read (7 families) to 656,093 reads (miR167), indicating that expression varies significantly among different miRNA families. This relative abundance is also observed in certain members from the same family. For example, the abundance of miR167 varied from 98 to 616,862 reads, as was the case for some other families, such as miR166 (1 to 381,733 reads), miR159 (2 to 235,279 reads) and miR396 (2 to 217,485 reads). These results indicate that different members have variable expression levels within one miRNA family.

Since the genome of *E. uniflora* is not publically available, the small RNA library was matched against a set of *de novo* assembled contigs from the *E. uniflora* leaf RNA-seq to identify putative miRNA precursor sequences. Candidate sequences with hairpin-like structures and mature miRNAs anchored in either or both of the 5p or 3p arms were further considered (Figure S2). Initial analysis allowed for the identification of 25 precursor sequences grouped into 15 conserved families (Table 3). The average value of MFE was -66.51 in these precursors and included two precursors (MIR167-2 and MIR169) with extreme values due to their long sizes. With respect to the % GC and MFEI, the average values were 47.63 and −0.94, respectively.

Within the identified families, MIR167 was the most abundant, with 676,895 reads, and contained 4 members (MIR167-1, MIR167-2, MIR167-3 and MIR167-4). In addition, several miRNA isoforms were detected in the libraries, and several of these were more abundant than the known miRNAs reported in miRBase for other plants (Figure 3). Furthermore, in the family MIR397, one precursor was identified with a typical structure and mature reads in the sense and antisense orientations. Both orientations were considered two independent precursors from the same family for the following analysis.

Identification of Novel miRNAs in *E. uniflora*

Using the previously described criteria in the identification of conserved pre-miRNAs, we obtained another 17 potential miRNA candidates grouped into 15 families (Table 4). In addition to the hairpin structure, the detection of miRNA* in 14 precursors is a strong indication to consider these miRNAs as true candidates. Comparisons among the mature sequences of candidate miRNAs and those miRNAs deposited in miRBase suggest that these candidates are novel miRNAs that have not been identified in others species and are possibly specific to the Myrtaceae family. These novel miRNAs displayed an average negative folding value of −137.89, which included 4 miRNAs with long sizes similarly observed in some previously identified conserved pre-miRNAs. With respect to the % GC and MFEI, the average values were 42.86 and −1.05, respectively. In addition, one novel pre-miRNA was found with mature sequences in the sense and antisense orientations and was considered to represent 2 members of the same family (nMIR001-1 and −2).

Biological Confirmation of Identified miRNAs in *E. uniflora*

The stem-loop RT-PCR method was used to validate the expression of seventeen conserved miRNAs (eun-MIR156, eun-MIR159, eun-MIR160, eun-MIR166, eun-MIR167-1, eun-MIR167-2, eun-MIR167-3, eun-MIR167-4, eun-MIR395, eun-MIR396-1, eun-MIR396-2, eun-MIR397-1, eun-MIR397-2, eun-MIR482-1, eun-MIR482-2, eun-MIR530, eun-MIR827) and ten novel miRNAs (eun-MIR001-1, eun-MIR001-2, eun-MIR004-2, eun-MIR005, eun-MIR006, eun-MIR008, eun-MIR009, eun-MIR012, eun-MIR013, eun-MIR014). We confirmed that these miRNAs were expressed in three different individuals collected *in situ* (Figure S3).

Identification and Classification of miRNA Targets

To understand the biological function of miRNAs in *E. uniflora*, the putative mRNA target sites of miRNA candidates were identified by aligning the most abundant mature miRNAs of each conserved and novel precursor to a set of *E. uniflora* assembled unigenes using psRNA target with default parameters and a maximum expectation value of 4. We found 87 potential targets in total, where 52 were targets of conserved miRNAs and 35 were targets of novel miRNAs, with an approximate average of 3 targets per miRNA. Detailed annotation results are given in Table 5 and S3.

Among the most important miRNA targets, also previously identified in other plants, we found the squamosa promoter binding protein (SBP)-like (SPL) genes, which are targets of the miR156 family and have functions that are conserved across plant species [17], affecting diverse developmental processes, such as leaf development, shoot maturation, phase change and flowering in plants [35-40]. We also identified the auxin response factor (ARF), a plant-specific family of DNA binding proteins involved in

Table 3. Pre-miRNAs identified in *E. uniflora* with sequence similarities to plant conserved miRNA families.

miRNA	Precursor miRNA						Mature miRNA				
	Contig code	Length	GC %	MFE	AMFE	MFEI	5p sequence more abundant	Read count	3p more abundant sequence	Read count	Total reads
eun-MIR156	Contig92889	97	54.64	−55.20	−56.91	−1.04	TTGACAGAAGATAGAGAGCAC	83933	GCTCTCCCTCTCCTGTCAACA	1	85396
eun-MIR159	Contig93245	163	45.40	−55.86	−34.27	−0.75	AGCTGCTGGTCTATGGATCCC	376	CTTGCATATGCCAGGAGCTTC	493	1302
eun-MIR160	Contig81816	116	53.45	−54.30	−46.81	−0.88	TGCCTGGCTCCCTGTATGCCA	293	GCGTATGAGGAGCCAAGCATA	22	323
eun-MIR162	Contig165400	108	48.15	−35.80	−33.15	−0.69	GGAGGCAGCGGTTCATCGATC	24	TCGATAAACCTCTGCATCCAG	10713	10884
eun-MIR166	Contig94223	228	43.42	−72.40	−31.75	−0.73	GGAATGTTGTCTGGCTCGAGG	11016	TCGGACCAGGCTTCATTCCCC	381733	409546
eun-MIR167-1	Contig126350	90	45.56	−50.00	−55.56	−1.22	TGAAGCTGCCAGCATGATCTGA	616862	AGATCATCTGGCAGTTTCAAC	262	620306
eun-MIR167-2	Contig163487	615	37.40	−163.60	−26.60	−0.71	TGAAGCTGCCAGCATGATCTGG	32188	TCAGGTCATCTTGCAGCTTCA	939	34461
eun-MIR167-3	Contig784s	81	48.15	−38.60	−47.65	−0.99	TGAAGCTGCCAGCGTGATCTCA	16305	ATCAGATCATGTGGCAGCTTCACC	73	22056
eun-MIR1674	Contig784a	81	48.15	−38.70	−47.78	−0.99	TGAAGCTGCCACATGATCTGA	71	ND	–	72
eun-MIR169	Contig142088	730	39.18	−160.80	−22.03	−0.56	TTATAGGCGATTGGAGGTATG	876	TTAGCTAAAGTCGTCTTGCCCA	6818	8961
eun-MIR172-1	Contig83802	122	47.54	−55.10	−45.16	−0.95	CAGGTGTAGCATCATCAAGAT	36	AGAATCTTGATGATGCTGCAT	495	1076
eun-MIR172-2	Contig113567	92	42.39	−39.70	−43.15	−1.02	GCAGCATCATCAAGATTCACA	12	AGAATCTTGATGATGCTGCAT	495	523
eun-MIR172-3	Contig85928	165	43.64	−71.20	−43.15	−0.99	GTAGCATCATCAAGATTCACA	33	AGAATCTTGATGATGCTGCAT	495	1048
eun-MIR395	Contig114717	88	51.14	−47.50	−53.98	−1.06	TCCCCTAGAGTTCTCCTGAACA	107	ATGAAGTGTTTGGGGGAACTC	1356	1659
eun-MIR396-1	Contig94388	160	42.50	−68.60	−42.88	−1.01	TTCCACGGCTTTCTTGAACTG	217485	GTTCAATAAAGCTGTGGGAAG	2028	223828
eun-MIR396-2	Contig153308	128	42.19	−49.10	−38.36	−0.91	TTCCACAGCTTTCTTGAACTG	23061	GTTCAAGCTAGCTGTGGGAAG	12981	64439
eun-MIR397-1	Contig87345s	126	51.59	−68.80	−54.60	−1.06	TCATTGAGTGCAGCGTTGAT	626	CGGTTTCGACAGCGCTGCACT	59	1052
eun-MIR397-2	Contig87345a	126	51.59	−63.20	−50.16	−0.97	TGCAGCGCTGTCGAAACCGAT	20	TCAACGCTGCACTCAATGATG	273	322
eun-MIR482-1	Contig88445	153	53.90	−94.60	−61.43	−1.14	CATGGGTTGTTTGGTGAGAGG	24202	TCTTGCCAATACCACCCATGCC	70833	100235
eun-MIR482-2	Contig88445	139	53.96	−71.00	−51.08	−0.95	GAAATGGGAGGGTGGGAAAGA	982	TTTCCTATTCCTCCCATTCCAT	3371	5574
eun-MIR482-3	Contig85065	169	50.89	−92.40	−54.67	−1.07	GAGATTCGAGCTACCGGAAGTTGTG	329	TTCCCAAGGCCGCCCATTCCGA	14915	17039
eun-MIR530	Contig18750	183	51.37	−82.30	−44.97	−0.88	TCTGCATTTGCACCTGCACCT	185	AGGTGCGGGTGCAGGTGCAGA	12	280
eun-MIR535-1	Contig68094	102	50.00	−42.60	−41.76	−0.84	TGACAACGAGAGAGAGCACGC	62562	GTGCTCTCTATCGCTGTCATA	4199	76334
eun-MIR535-2	Contig71803	100	49.00	−47.00	−47.00	−0.96	TGACAACGAGAGAGAGCACGC	62562	TGCTCTCTACCGTTGTCATG	116	72263
eun-MIR827	Contig93928	81	45.68	−44.30	−54.69	−1.20	CTTTGTTGATGGCCATCTAATC	27	TTAGATGACCATCAGCGAACA	266	304

MFE: minimal folding free energy (kcal/mol); AMFE: Adjusted MFE; MFEI: minimal folding free energy index; ND: no detected.

doi:10.1371/journal.pone.0049811.t003

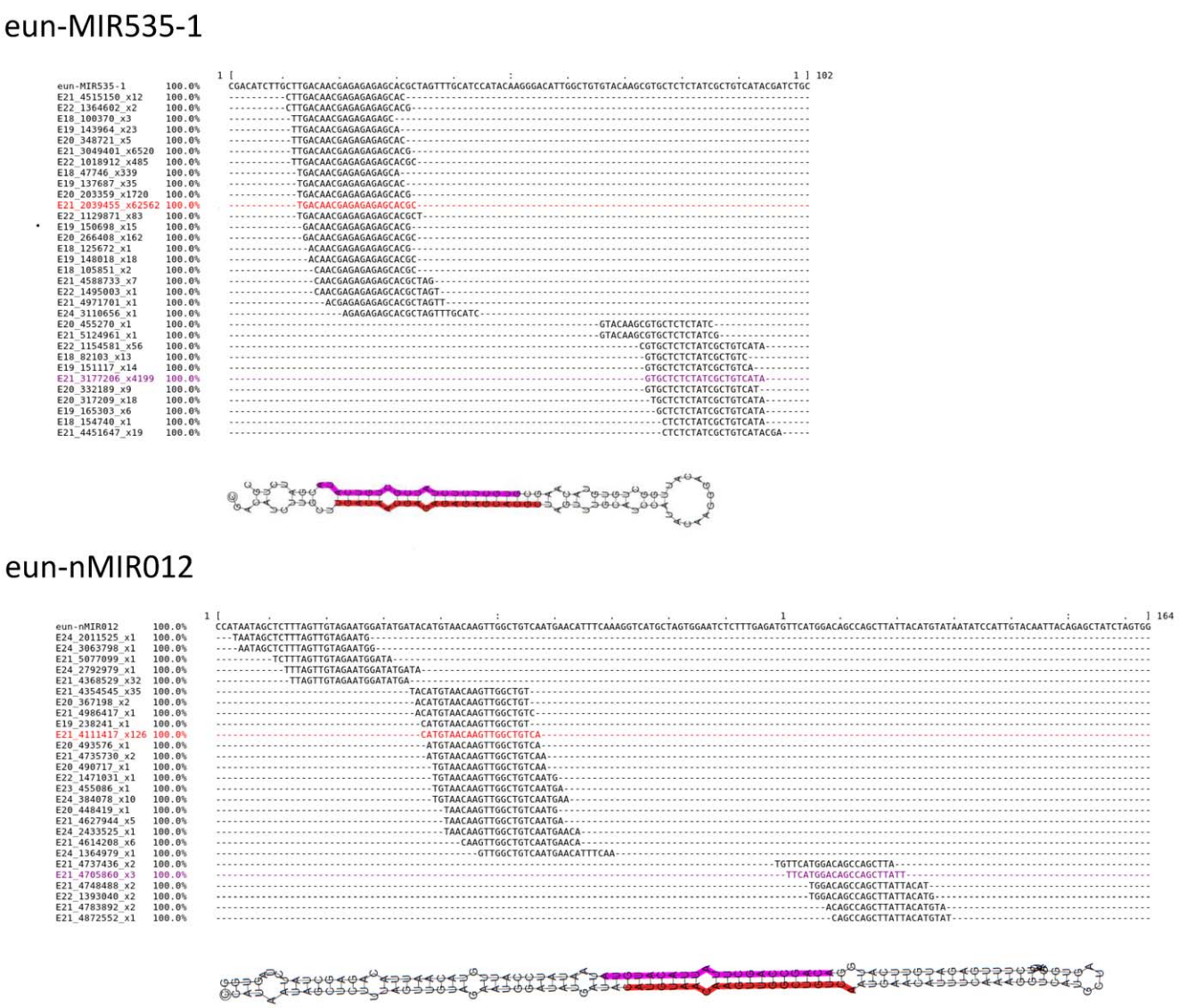

Figure 3. Predicted secondary structures of conserved and novel miRNAs of *E. uniflora*. Secondary structures of the precursors eun-MIR535-1 and eun-nMIR012, their locations and the expression of small RNAs mapped onto these precursors. Sequences of the most abundant mature miRNAs in the 5p and 3p arms are labeled in purple and red, respectively. Values on the left side of the miRNA sequence represent read counts in the leaf library.
doi:10.1371/journal.pone.0049811.g003

hormone signal transduction that are targets for the miRNA families miR167 and miR160 [15,41,42]. Another important gene identified and targeted by miR162, with a significant role in the regulation of gene expression, is the pentatricopeptide repeat gene (PPR). This gene belongs to a large family implicated in post-transcriptional processes, such as splicing, editing, processing and translation specifically in organelles like mitochondria and chloroplasts [43]. These results substantiate the in silico identification of conserved and novel targets from *E. uniflora*.

All targets regulated by the conserved and novel miRNAs identified in this study were subjected to GO analysis to evaluate their potential functions. The categorization of these genes, according to biological processes, cellular components and molecular functions, is summarized in Figure 4. Based on biological processes, these targets were classified into 13 categories, and the three most overrepresented GO terms, either for conserved or novel miRNAs, were cellular processes, metabolic processes and responses to stimulus, suggesting that *Eugenia* miRNAs are involved in a broad range of physiological functions. Categories based on molecular function revealed that the target genes were related to 7 functions, and the four most frequent terms were protein binding, nucleotide binding, hydrolase activity and nucleic acid binding. In the category of cellular components, the analysis revealed that the protein products from the genes targeted by conserved and novel miRNAs are expressed mainly in the plastid and nucleus.

The iPATH2 server was used to produce an overview of the metabolic pathways involved in the secondary metabolites synthesis and potentially regulated by miRNAs in *E. uniflora*. Our results showed that three enzymes involved in several types of metabolism and secondary metabolites are regulated by identified miRNAs (Figure S4). The phosphoglycerate mutase is a potential target of eun-MIR396-2 and is involved in the pathway of gluconeogenesis while the hydroxyphenylpyruvate reductase is

Table 4. New putative miRNA precursors identified in *E. uniflora*.

miRNA	Precursor miRNA						Mature miRNA				
	Contig name	Length	GC %	MFE	AMFE	MFEI	5p more abundant sequence	Read count	3p more abundant sequence	Read count	Total reads
eun-nMIR001-1	Contig164780s	820	54.39	−451.70	−55.09	−1.01	TCGGCTGTCAATTTCTGGATT	185919	ATCCAGAAATTGGCAGCCGTT	110	192103
eun-nMIR001-2	Contig164780a	820	54.39	−446.00	−54.39	−1.00	CAATTTCTGGATTTCAGTTCG	20	TCGAACTGAAATCCAGAAATT	2	63
eun-nMIR002	Contig29785	173	30.06	−56.70	−32.77	−1.09	CGAAAAATGATTGGTTGTATCGCT	18	CGATCTAATCAATCATTTTTCGGG	2	21
eun-nMIR003	Contig121100	110	48.18	−63.20	−57.45	−1.19	TACTCGTTCCGTTGATCCATC	88	TGGATCAATAGAACGAGCAGGTGA	157	561
eun-nMIR004-1	Contig37387s	104	29.81	−47.10	−45.29	−1.52	TCGTAAATCCACTATATCTCT	3	TAGATATAGTGGATTTTCGAT	9	13
eun-nMIR004-2	Contig37387a	104	29.81	−46.40	−44.62	−1.50	ND	–	CAGAGATATAGTGGATTTACG	314	385
eun-nMIR005	Contig143563	106	34.91	−21.40	−20.19	−0.58	ND	–	GAGAATGATGAGTTAAATGGA	12	30
eun-nMIR006	Contig113617	113	33.63	−42.70	−37.79	−1.12	TCTCTGTTGATCTGATAAATA	19	TTTGTCGGATGAACAGGGAAT	18	55
eun-nMIR007	Contig116664r	96	46.88	−55.80	−58.13	−1.24	TAGGGTCAGATCGCTACTTAG	211	TAAGTGGTGATCTGACTCTAA	4446	5750
eun-nMIR008	Contig79716	426	43.66	−228.00	−53.52	−1.23	TCGAGCCCCTCCCACAGATTG	313	ATCTGTGGAAGAGACTCGACT	8403	13617
eun-nMIR009	Contig81320	1545	37.41	−397.90	−25.75	−0.69	TTCAAGTCTAACAACCTCAGCT	5774	TGGAGGTTGTTTGGCTTGAGCT	1968	11577
eun-nMIR010	Contig84248	350	48.86	−143.30	−40.94	−0.84	TGCTGTTCTTCCGTTCACGAAT	309	TCGTGAAGGAAGAATGTGCAAT	6340	10307
eun-nMIR011	Contig164559	207	51.21	−98.70	−47.68	−0.93	GCTCGAGGTCAGTTTGTCGCC	1553	CGGCAAACTGGACCTCGAGATC	110	2884
eun-nMIR012	Contig167957	164	35.98	−91.80	−55.98	−1.56	CATGTAACAAGTTGGCTGTCA	126	TTCATGGACAGCCAGCTTATT	3	243
eun-nMIR013	Contig87665	179	56.42	−84.90	−47.43	−0.84	TGAAGCAGATCAAGAACCCAG	155	TCTCGTTCCGCTTCATCTGAA	2	172
eun-nMIR014	Contig200629r	87	55.17	−23.40	−26.90	−0.49	ND	–	GCATCACTAGCTTACGCTCTG	30	159
eun-nMIR015	Contig165886	124	37.90	−45.10	−36.37	−0.96	ND	–	CAATGAACGCATTTGCAGGTG	21	22

MFE: minimal folding free energy (kcal/mol); AMFE: Adjusted MFE; MFEI: minimal folding free energy index; ND: no detected.

doi:10.1371/journal.pone.0049811.t004

Table 5. Predicted targets of novel miRNAs in *E. uniflora*.

miRNA	Inhibition	Score*	Putative Function
eun-nMIR001	Cleavage	1.5	Atp-dependent helicase rhp16-like
	Cleavage	3	Cytochrome p450
	Cleavage	3.5	Long chain acyl- synthetase 9
	Cleavage	3.5	Kinesin-related protein
	Translation	3	Transcription initiation factor iib
	Translation	3	Brassinazole-resistant 1
	Translation	3.5	Myosin family protein with dil domain
eun-nMIR002	Cleavage	3	Serine threonine-protein phosphatase 2a regulatory subunit b subunit alpha-like
	Cleavage	3.5	Probable receptor-like protein kinase at1g67000-like
eun-nMIR003	Cleavage	4	Udp-glycosyltransferase 74b1
eun-nMIR004	Cleavage	2.5	Auxin efflux carrier protein
	Cleavage	3.5	Sucrose nonfermenting 4-like
	Translation	3.5	Type i inositol- -trisphosphate 5-phosphatase 2-like
eun-nMIR005	Translation	3	Pentatricopeptide repeat-containing protein
	Cleavage	3.5	Protein reticulata-related 1
	Cleavage	3.5	Agenet domain-containing protein
eun-nMIR006	Translation	3.5	Outward rectifying potassium channel
eun-nMIR007	Cleavage	3.5	Aspartate semialdehyde
	Cleavage	3	Primary-amine oxidase
eun-nMIR008	Translation	3.5	Adenosine deaminase
eun-nMIR009	Cleavage	3	Cc-nbs-lrr resistance protein
eun-nMIR010	Translation	4	P8mtcp1
	Cleavage	4	Nbs-lrr resistance protein
	Translation	4	Cullin-1-like isoform 1
eun-nMIR011	Translation	3	E3 ubiquitin-protein ligase upl7
	Cleavage	3.5	Eukaryotic peptide chain release factor subunit 1-1
eun-nMIR012	Translation	3.5	Tho complex subunit 2
eun-nMIR013	Translation	3.5	Beta-amylase
	Translation	3.5	Integral membrane single c2 domain protein
eun-nMIR014	Cleavage	0	Ycf68 protein
eun-nMIR015	Cleavage	3.5	Rna-binding motif x-linked 2
	Cleavage	3.5	Transducin wd-40 repeat-containing protein
	Cleavage	3.5	Clip-associating protein
	Cleavage	3.5	Probable exocyst complex component 4-like
	Cleavage	3.5	Photosystem i p700 apoprotein a1

*psRNATarget value.

doi:10.1371/journal.pone.0049811.t005

targeted by eun-MIR162 and is involved in terpenoid-quinone, tropane, pireridine and pyridine biosynthesis. In a similar way, eun-nMIR007 regulates primary-amine oxidase, an enzyme involved in the tropane, pireridine, pyridine and isoquinoline alkaloid biosynthesis.

Discussion

Though several miRNAs have been identified via computational or experimental approaches in different plant families, there is no sequence or functional information available about miRNAs in any Myrtaceae species, which are economically important in the spice, fruit, timber and pharmacology industries [44].

We used Solexa technology for deep sequencing of a small RNA library to identify miRNAs in *E. uniflora*. The length distribution pattern obtained indicates that the majority of the redundant small RNAs from the library were 21 nt in length, which is atypical because 24 nt is the most abundant size produced by DCL3 in other plants [45]. This distribution pattern is similar to those observed in previous reports of plant small RNA sequencing using Solexa technology, such as wheat [46], grapevine [47], melon [48] and trifoliate orange [49], suggesting that the composition of the small RNA population varies among species. Additionally, other important causes for this variation include the developmental stage and environmental conditions in which the sample was collected. Contrary to the results observed with the redundant sequences, the analysis of the unique sequences showed that 24 nt was the

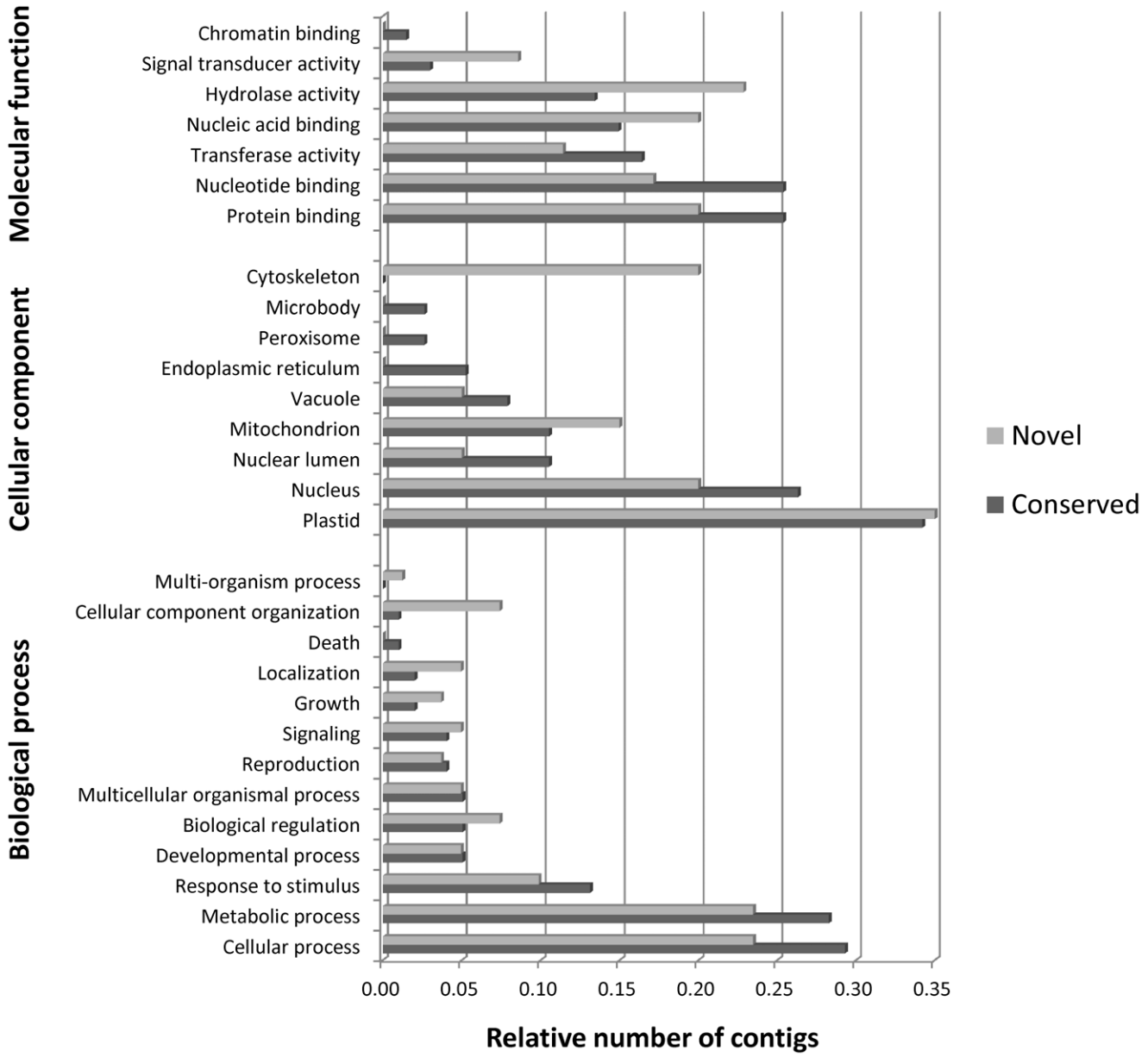

Figure 4. Gene categories and the distribution of target genes of the most abundant mature miRNAs in the conserved and novel pre-miRNA identified in *E. uniflora*.
doi:10.1371/journal.pone.0049811.g004

dominant read length in comparison to all other sequence lengths, and similar results have been observed in other studies [50-53]. Small RNAs of 24 nt in length are known to be involved in heterochromatin transcriptional silencing in genomes with a high content of repetitive sequences [54], indicating the possible genome complexity of *E. uniflora*.

In this study, we compared our small RNA library from *E. uniflora* against known plant miRNAs from the miRBase database and identified 204 conserved miRNAs from different species grouped into 45 families. High throughput sequencing, which has the ability to generate millions of small RNA sequences, is a powerful tool to estimate expression profiles of miRNA. This technology provides the resources to determine the abundance of various miRNA families and even distinguish among different members of a given family. In our case, we found significant differences among the number and abundance of the members identified in each family, which is in agreement with previous studies [48,55,56] and suggests that this wide variation is due to a functional divergence in the conserved miRNA families.

Although conserved miRNAs have been identified by sequencing and comparison against miRNAs from other species, most plant species-specific miRNAs remain unidentified due to their lower levels of expression, which result in a small number of sequenced reads in comparison to the conserved miRNAs [57]. For this reason, we used a new approach to identify novel miRNAs in species where genomic data and resources were not available. We made use of simultaneous sequence comparison of small RNA and RNAseq libraries. Using this methodology, we identified 17 potential miRNA candidates specific for *E. uniflora*. From these, 14 contained complementary antisense miRNA, which provided

more evidence for their existence as novel miRNAs, as observed in cucumber [51] and grapevine [53]. The other miRNAs that do not satisfy this last criterion require further investigation for their confirmation as miRNAs.

To understand the function of the identified miRNAs, their putative targets were predicted using a bioinformatics approach. Several identified targets of conserved miRNAs of *E. uniflora* are transcriptional factors, similar to the results reported in other studies [47,48,58,59]. In the case of the novel miRNA targets, we found that the transcription initiation factor iib and the pentatricopeptide repeat-containing proteins are targeted by eun-nMIR001 and eun-nMIR005, respectively.

It has been reported in *Arabidopsis* that regulation of *ARF17* by miR160 is important for growth and development [60], regulation of *ARF6* and *8* by miR167 is important for development of anthers and ovules [61] and regulation of *ARF10* and *16* by miR160 plays a role in root cap formation [62]. In the present study, we found that *ARF17* is regulated by eun-MIR160 while other members of ARF family were not targeted by eun-MIR167. This discrepancy agrees with the previously reported in *Arabidopsis* because we used a leaf transcriptome as reference for the target identification. We confirm this observation not found homologs for *AtARF6*, 8, *10* and *16* by BLASTx in *E. uniflora* transcriptome.

In addition, with the analysis of GO terms, we identified 3 candidate targets likely involved in the response to abiotic stress: ATP-dependent helicase rhp16-like (eun-nMIR002), sucrose nonfermenting 4-like (eun-nMIR004) and serine threonine-protein phosphatase 2a regulatory subunit b'' subunit alpha-like (eun-nMIR002). The sucrose nonfermenting 4-like (SNF4) protein is a subunit of the probable trimeric SNF1-related protein kinase (SnRK) complex, which may play a role in a signal transduction cascade regulating gene expression and carbohydrate metabolism in higher plants [63]. Otherwise, the serine threonine-protein phosphatase 2A regulatory subunit b ''subunit alpha-like PP2Ab'' is a structural subunit of the Ser/Thr phosphatases holoenzyme (PP2A) and recent studies suggesting the possible physiological role of PP2A in the drought stress response [64]. These results indicated that the targets from novel miRNAs identified here are possibly related to the adaptation of *E. uniflora* to different types of stress and environmental conditions observed *in natura*. Future experimental validation will determine how many of these predicted targets are genuinely targeted by miRNAs in specific environmental and physiological conditions.

Interestingly, we found three miRNAs involved in the regulation of enzymes that play critical roles in secondary metabolites synthesis. These findings suggest that variation in the levels of expression of these miRNAs could alter the levels of production of certain types of secondary metabolites. It is consistent with the previous reports that the concentration of these metabolites varies between specimens of *E. uniflora* from different geographical locations [2,3]. More studies are necessary to confirm these preliminary findings and evaluate the correlation between the miRNA expression and secondary metabolite production.

Conclusions

In summary, this study provides the first view of the diversity of miRNAs and their abundance in Myrtaceae and strongly supports the idea that miRNAs play an important conserved role in several physiological processes, as previously proposed for other plants. Our bioinformatics analysis indicates that miRNAs might contribute to different processes by affecting multiple target genes and different signaling pathways. Although the exact function of these miRNA target genes remains to be confirmed, we believe the present study provides novel insights into the molecular processes involved in conserved miRNA function.

Supporting Information

Figure S1 Flow chart of procedures for miRNA identification.
(TIF)

Figure S2 Represetation of the predicted secondary structures of the conserved and novel miRNA precursors of *E. uniflora* and the locations of the more abundant mature miRNAs.
(PDF)

Figure S3 Detection of miRNA expression in different *E. uniflora* individuals by RT-PCR. Products generated by stem-loop RT-PCR were resolved on a 2% agarose. Leaf samples from three independent Eugenia uniflora trees were used to evaluate the presence of each miRNA.
(TIFF)

Figure S4 iPath secondary metabolite map showing the different pathways where are involved each evaluated enzyme. Each grey dot represents a metabolite and each colored line represents the different route affected by the enzyme targeted. In red: phosphoglycerate mutase (regulated by eun-MIR396-2). In blue: primary-amine oxidase (regulated by eun-nMIR007).
(TIFF)

Table S1 Length distribution and read sequence abundance in the *E. uniflora* small RNA library.
(XLS)

Table S2 Abundance of predicted plant conserved miRNAs in the small RNA library of *E. uniflora*.
(XLS)

Table S3 Predicted transcript targets of plant conserved miRNAs in *E. uniflora*.
(XLS)

Acknowledgments

The authors would like to thank Maria Elena Gonzalez for reviewing of the manuscript.

Author Contributions

Conceived and designed the experiments: RM. Performed the experiments: FG MPA RM. Analyzed the data: FG MPA APK GLM RM. Contributed reagents/materials/analysis tools: FG MPA APK GLM RM. Wrote the paper: FG. Reviewed the manuscript: MPA APK GLM RM.

References

1. Burt S (2004) Essential oils: their antibacterial properties and potential applications in foods–a review. International journal of food microbiology 94: 223–253.
2. Lago JHG, Souza ED, Mariane B, Pascon R, Vallim M a, et al. (2011) Chemical and biological evaluation of essential oils from two species of Myrtaceae - *Eugenia uniflora* L. and *Plinia trunciflora* (O. Berg) Kausel. Molecules 16: 9827–9837.
3. Victoria FN, Lenardão EJ, Savegnago L, Perin G, Jacob RG, et al. (2012) Essential oil of the leaves of *Eugenia uniflora* L.: antioxidant and antimicrobial properties. Food and chemical toxicology 50: 2668–2674.
4. Arai I, Amagaya S, Komatsu Y, Okada M, Hayashi T, et al. (1999) Improving effects of the extracts from *Eugenia uniflora* on hyperglycemia and hypertriglyceridemia in mice. Journal of Ethnopharmacology 68: 307–314.

5. Consolini a E, Baldini O a, Amat a G (1999) Pharmacological basis for the empirical use of *Eugenia uniflora* L. (Myrtaceae) as antihypertensive. Journal of Ethnopharmacology 66: 33–39.
6. Matsumura T, Kasai M, Hayashi T, Arisawa M, Momose Y, et al. (2000) α-glucosidase inhibitors from Paraguayan natural medicine, ñangapiry, the leaves of *Eugenia uniflora*. Pharmaceutical Biology 38: 302–307.
7. Ogunwande I, Olawore N, Ekundayo O, Walker T, Schmidt J, et al. (2005) Studies on the essential oils composition, antibacterial and cytotoxicity of *Eugenia uniflora* L. International Journal of Aromatherapy 15: 147–152.
8. Luzia DM, Bertanha BJ, Jorge N (2010) Pitanga (*Eugenia uniflora* L.) seeds?: antioxidant potential and fatty acids profile. Revista do Instituto Adolfo Luz 69: 175–180.
9. Santos KK a, Matias EFF, Tintino SR, Souza CES, Braga MFBM, et al. (2012) Anti-Trypanosoma cruzi and cytotoxic activities of *Eugenia uniflora* L. Experimental Parasitology.
10. Almeida DJ, Faria MV, Da Silva PR (2012) Biologia experimental em Pitangueira: uma revisão de cinco décadas de publicações científicas. Ambiência Guarapuava (PR) 8: 159–175.
11. Mallory AC, Vaucheret H (2006) Functions of microRNAs and related small RNAs in plants. Nature Genetics 38 Suppl: S31–6.
12. Denli AM, Tops BBJ, Plasterk RH a, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. Nature 432: 231–235.
13. Bushati N, Cohen SM (2007) MicroRNA functions. Annual Review of Cell and Developmental Biology 23: 175–205.
14. Axtell MJ, Westholm JO, Lai EC (2011) Vive la différence: biogenesis and evolution of microRNAs in plants and animals: 1–13.
15. Sunkar R, Li Y-F, Jagadeeswaran G (2012) Functions of microRNAs in plant stress responses. Trends in Plant Science 17: 196–203.
16. Dezulian T, Palatnik JF, Huson D, Weigel D (2005) Conservation and divergence of microRNA families in plants. Bioinformatics 6: P13.
17. Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAS and their regulatory roles in plants. Annual Review of Plant Biology 57: 19–53.
18. Allen E, Xie Z, Gustafson AM, Sung G-H, Spatafora JW, et al. (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nature Genetics 36: 1282–1290.
19. Gambino G, Perrone I, Gribaudo I (2008) A Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochemical Analysis 19: 520–525.
20. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864.
21. Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, et al. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Research 37: D159–62.
22. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35: 7188–7196.
23. He S, Liu C, Skogerbø G, Zhao H, Wang J, et al. (2008) NONCODE v2.0: decoding the non-coding. Nucleic Acids Research 36: D170–2.
24. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.
25. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.
26. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, et al. (2010) Tablet–next generation sequence assembly visualization. Bioinformatics 26: 401–402.
27. Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, et al. (2008) A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics 24: 2252–2253.
28. Chen C, Ridzon D a, Broomer AJ, Zhou Z, Lee DH, et al. (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic acids research 33: e179.
29. Kulcheski FR, de Oliveira LF, Molina LG, Almerão MP, Rodrigues F, et al. (2011) Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC genomics 12: 307.
30. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, et al. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651–652.
31. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome research 9: 868–877.
32. Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Research 39: W155–9.
33. Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics 2008: 619832.
34. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) iPath2.0: interactive pathway explorer. Nucleic acids research 39: W412–5.
35. Poethig RS (2009) Small RNAs and developmental timing in plants. Current Opinion in Genetics & Development 19: 374–378.
36. Wu G, Park MY, Conway SR, Wang J-W, Weigel D, et al. (2009) The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. Cell 138: 750–759.
37. Schwarz S, Grande AV, Bujdoso N, Saedler H, Huijser P (2008) The microRNA regulated SBP-box genes SPL9 and SPL15 control shoot maturation in *Arabidopsis*. Plant Molecular Biology 67: 183–195.
38. Shikata M, Koyama T, Mitsuda N, Ohme-Takagi M (2009) Arabidopsis SBP-box genes SPL10, SPL11 and SPL2 control morphological change in association with shoot maturation in the reproductive phase. Plant & Cell Physiology 50: 2133–2145.
39. Xie K, Wu C, Xiong L (2006) Genomic organization, differential expression, and interaction of SQUAMOSA promoter-binding-like transcription factors and microRNA156 in rice. Plant Physiology 142: 280–293.
40. Chuck G, Cigan a M, Saeteurn K, Hake S (2007) The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA. Nature Genetics 39: 544–549.
41. Wu M-F, Tian Q, Reed JW (2006) Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. Development 133: 4211–4218.
42. Yang JH, Han SJ, Yoon EK, Lee WS (2006) "Evidence of an auxin signal pathway, microRNA167-ARF8-GH3, and its response to exogenous auxin in cultured rice cells." Nucleic Acids Research 34: 1892–1899.
43. Fujii S, Small I (2011) The evolution of RNA editing and pentatricopeptide repeat genes. The New phytologist 191: 37–47.
44. Okoh-Esene Rosemary, Husseini Suleiman AT (2011) Proximate and phyto-chemical analysis of leaf, stem and root of *Eugenia uniflora* (Surinam or Pitanga cherry). Journal of Natural Product and Plant Resources 1: 1–4.
45. Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, et al. (2004) Genetic and functional diversification of small RNA pathways in plants. PLoS Biology 2: E104.
46. Yao Y, Guo G, Ni Z, Sunkar R, Du J, et al. (2007) Cloning and characterization of microRNAs from wheat (*Triticum aestivum* L.). Genome Biology 8: R96.
47. Pantaleo V, Szittya G, Moxon S, Miozzi L, Moulton V, et al. (2010) Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. The Plant Journal 62: 960–976.
48. Gonzalez-Ibeas D, Blanca J, Donaire L, Saladié M, Mascarell-Creus A, et al. (2011) Analysis of the melon (*Cucumis melo*) small RNAome by high-throughput pyrosequencing. BMC Genomics 12: 393.
49. Zhang J-Z, Ai X-Y, Guo W-W, Peng S-A, Deng X-X, et al. (2011) Identification of miRNAs and Their Target Genes Using Deep Sequencing and Degradome Analysis in Trifoliate Orange [*Poncirus trifoliate* (L.) Raf]. Molecular Biotechnology: 44–57.
50. Song C, Wang C, Zhang C, Korir NK, Yu H, et al. (2010) Deep sequencing discovery of novel and conserved microRNAs in trifoliate orange (*Citrus trifoliata*). BMC Genomics 11: 431.
51. Martínez G, Forment J, Llave C, Pallás V, Gómez G (2011) High-throughput sequencing, characterization and detection of new and conserved cucumber miRNAs. PloS One 6: e19523.
52. Schreiber AW, Shi B-J, Huang C-Y, Langridge P, Baumann U (2011) Discovery of barley miRNAs through deep sequencing of short reads. BMC Genomics 12: 129.
53. Wang C, Wang X, Kibet NK, Song C, Zhang C, et al. (2011) Deep sequencing of grapevine flower and berry short RNA library for discovery of novel microRNAs and validation of precise sequences of grapevine microRNAs deposited in miRBase. Physiologia Plantarum 143: 64–81.
54. Simon S a, Meyers BC (2011) Small RNA-mediated epigenetic modifications in plants. Current Opinion in Plant Biology 14: 148–155.
55. Zhao C-Z, Xia H, Frazier TP, Yao Y-Y, Bi Y-P, et al. (2010) Deep sequencing identifies novel and conserved microRNAs in peanuts (*Arachis hypogaea* L.). BMC Plant Biology 10: 3.
56. Puzey JR, Karger A, Axtell M, Kramer EM (2012) Deep Annotation of *Populus trichocarpa* microRNAs from Diverse Tissue Sets. PloS One 7: e33034.
57. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, et al. (2007) High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PloS One 2: e219.
58. Colaiacovo M, Subacchi A, Bagnaresi P, Lamontanara A, Cattivelli L, et al. (2010) A computational-based update on microRNAs and their targets in barley (*Hordeum vulgare* L.). BMC Genomics 11: 595.
59. Lv S, Nie X, Wang L, Du X, Biradar SS, et al. (2012) Identification and Characterization of MicroRNAs from Barley (*Hordeum vulgare* L.) by High-Throughput Sequencing. International Journal of Molecular Sciences 13: 2973–2984.
60. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, et al. (2002) Prediction of plant microRNA targets. Cell 110: 513–520.
61. Wu M-F, Tian Q, Reed JW (2006) *Arabidopsis* microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. Development 133: 4211–4218.
62. Wang J, Wang L, Mao Y, Cai W, Xue H, et al. (2005) Control of Root Cap Formation by MicroRNA-Targeted Auxin Response Factors in *Arabidopsis*. The Plant Cell 17: 2204–2216.
63. Kleinow T, Bhalerao R, Breuer F, Umeda M, Salchert K, et al. (2000) Functional identification of an *Arabidopsis* snf4 ortholog by screening for heterologous multicopy suppressors of snf4 deficiency in yeast. 23: 115–122.
64. Xu C, Jing R, Mao X, Jia X, Chang × (2007) A wheat (*Triticum aestivum*) protein phosphatase 2A catalytic subunit gene provides enhanced drought tolerance in tobacco. Annals of botany 99: 439–450.

Capítulo 4

De novo assembly of *Eugenia uniflora* L. transcriptome and identification of genes from the terpenoid biosynthesis pathway

Frank Guzman[a,b], Franceli Rodrigues Kulcheski[a,b], Andreia Carina Turchetto-Zolet[a], Rogerio Margis[a,b,c]

a) PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil

b) PPGBCM, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil

c) Departamento de Biofísica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil

Artigo publicado na Plant Science (2014)

Plant Science 229 (2014) 238–246

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

De novo assembly of *Eugenia uniflora* L. transcriptome and identification of genes from the terpenoid biosynthesis pathway

Frank Guzman[a,b], Franceli Rodrigues Kulcheski[a,b], Andreia Carina Turchetto-Zolet[a], Rogerio Margis[a,b,c,*]

[a] PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil
[b] PPGBCM, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil
[c] Departamento de Biofisica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 14 August 2014
Received in revised form 7 October 2014
Accepted 10 October 2014
Available online 22 October 2014

Keywords:
Eugenia uniflora
Myrtaceae
Transcriptome
Terpene synthase
Oxidosqualene cyclase
Secondary metabolism

ABSTRACT

Pitanga (*Eugenia uniflora* L.) is a member of the Myrtaceae family and is of particular interest due to its medicinal properties that are attributed to specialized metabolites with known biological activities. Among these molecules, terpenoids are the most abundant in essential oils that are found in the leaves and represent compounds with potential pharmacological benefits. The terpene diversity observed in Myrtaceae is determined by the activity of different members of the terpene synthase and oxidosqualene cyclase families. Therefore, the aim of this study was to perform a *de novo* assembly of transcripts from *E. uniflora* leaves and to annotation to identify the genes potentially involved in the terpenoid biosynthesis pathway and terpene diversity. In total, 72,742 unigenes with a mean length of 1048 bp were identified. Of these, 43,631 and 36,289 were annotated with the NCBI non-redundant protein and Swiss-Prot databases, respectively. The gene ontology categorized the sequences into 53 functional groups. A metabolic pathway analysis with KEGG revealed 8,625 unigenes assigned to 141 metabolic pathways and 40 unigenes predicted to be associated with the biosynthesis of terpenoids. Furthermore, we identified four putative full-length terpene synthase genes involved in sesquiterpenes and monoterpenes biosynthesis, and three putative full-length oxidosqualene cyclase genes involved in the triterpenes biosynthesis. The expression of these genes was validated in different *E. uniflora* tissues.

1. Introduction

Eugenia uniflora L., commonly known as Brazilian cherry, pitanga or nangapiri, belongs to the Myrtaceae family [1]. *E. uniflora* L. is a shrubby tree with edible cherry-like fruits, is native to South America and occurs in different vegetation types and ecosystems, with high adaptability to different soil and climate conditions [2]. Its leaves have been used in folk medicine for the treatment of different diseases due to the presence of different compounds in its essential oils [3]. For this reason, *E. uniflora* has been the focus of several phytochemical studies in recent years. These studies reported some pharmacological properties that were reviewed by Lim [4] and that encompass antioxidant, antimicrobial, antihyperglycemic, antihypertriglycerimedic, hypotensive, vasorelaxant, antiviral, antinociceptive, hypothermic, central nervous system-related, diuretic, anti-inflammatory, antidiarrheal, muscle contractile, trypanocidal, antibiotic potentiating and toxicity activities.

Among the main compounds identified in the essential oils of *E. uniflora* leaves, terpenes are the most abundant. These compounds are one of the most important in the plant kingdom because they form such a large class of plant specialized metabolites and play a number of roles in the interaction between a plant and its environment [5]. The more abundant terpenes identified in *E. uniflora* correspond to different types of sesquiterpenes (C15) and monoterpenes (C10) [6–10]. The amount of terpene found in the essential oils of *E. uniflora* specimens from different regions varies depending on soil composition, sample collection season, altitude and the method used to extract the essential oils [11].

The terpene diversity reported in plant species is a consequence of the enzymatic activity of terpene synthases (TPS) and

Abbreviations: EC, enzyme commission number; GO, gene ontology; MCMC, Markov chain Monte Carlo; ORF, open reading frame; TPS, terpene synthase; OSC, oxidosqualene cyclase; wAIC, Akaike's information criterion.

* Corresponding author at: Centre of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Population, building 43431, Federal University of Rio Grande do Sul - UFRGS, P.O. Box 15005, CEP 91501-970, Porto Alegre, RS, Brazil.
Tel.: +55 51 33087766; fax: +55 5133087309.
E-mail addresses: rogerio.margis@gmail.com, rogerio.margis@ufrgs.br (R. Margis).

http://dx.doi.org/10.1016/j.plantsci.2014.10.003
0168-9452/

oxidosqualene cyclases (OSC), which catalyze the conversion of a few common substrates into thousands of terpene structures [12]. More specifically, the members of the TPS family are involved in the biosynthesis of monoterpenes, diterpenes and sesquiterpenes [13]. Several studies have identified TPS genes in plants with available complete genomic sequences. For example, 32, 69 and 120 putative TPS genes were reported in *Arabidosis thaliana*, *Vitis vinifera* and *Eucalyptus grandis*, respectively [14–16]. Genes from the monoterpene pathway were also identified in *Melaleuca alternifolia* [17]. In the other way, triterpenes are biosynthesized from 2,3-oxidosqualene by the OSC genes and so far known as triterpenes synthases [18,19]. It has been reported that 13 and 12 OSC genes are present in *A. thaliana* and *Oryza sativa* L. ssp. *japonica* cv Nipponbare genomes, respectively [20]. Thus, genomic and transcriptomic studies in different species are very important to provide a better understanding of the transcriptional regulation of many genes involved in metabolic and signaling pathways that cause variation in the presence and abundance of the terpenes in plants grown under different conditions and different geographical locations.

RNAseq, or mRNA deep sequencing, emerged as a powerful tool to identify whole transcripts and to help in transcriptome analyses due its high-throughput, accuracy and reproducibility [21]. In plants, this approach has accelerated the understanding of complex transcriptional patterns and has provided measurements of gene expression in different tissues or at different stages of development [22]. In the present work, we provide the first reference transcriptome for *E. uniflora*, the type species of the genera [23]. We provide a general annotation that will be helpful for further studies, but we focused our analyses on the genes involved in the terpenoid synthesis pathway.

2. Materials and methods

2.1. Plant material and RNA extraction

Total RNA for the library construction was isolated from mature leaves of an *E. uniflora* tree grown in an orchard at the Federal University of Rio Grande do Sul (Porto Alegre, Brazil). Mature leaves, young leaves and petal tissues were collected as biological triplicates for the RT-qPCR analysis. All tissues were immediately frozen in liquid nitrogen and stored at −80 °C until RNA extraction. For total RNA extraction, a CTAB-based method was employed [24]. Briefly, 900 μL of extraction buffer (2% CTAB, 2.0% PVP-40, 2 M NaCl, 100 mM Tris–HCl pH 8.0, 25 mM EDTA pH 8.0 and 2% of β-mercaptoethanol added just before use) were heated at 65 °C in a microcentrifuge tube. Each sample, approximately 150 mg, was ground in liquid nitrogen, mixed with the extraction buffer and incubated at 65 °C for 10 min. An equal volume of chloroform:isoamyl alcohol (24:1, v/v) was added, and the tube was centrifuged at 7000 × *g* for 20 min at 4 °C. The supernatant was recovered, and a second extraction with chloroform:isoamyl alcohol was performed. The supernatant was transferred to a new microcentrifuge tube, and one half of the volume of 4 M LiCl was added. The mixture was incubated at −20 °C for 30 min, and RNA was selectively pelleted by centrifugation at 16,000 × *g* for 30 min at 4 °C. The pellet was washed with 75% ethanol, dried and solubilized in DEPC–water. RNA integrity was evaluated by electrophoresis on a 1% agarose gel.

2.2. Library construction and deep sequencing

Total RNA (>10 μg) was sent to Fasteris SA (Plan-les-Ouates, Switzerland) for processing. One RNAseq library was constructed using the following successive steps: poly-A purification, cDNA synthesis using a poly-T primer shotgun method to generate inserts of 500 nt, 3P and 5P adapter ligations, pre-amplification and colony generation. Finally, the library was sequenced using the Illumina HiSeq2000 platform. The output data obtained after single-end sequencing included sequence tags of 100 bases and are available at NCBI Gene Expression Omnibus (GEO) with the accession number GSE38212.

2.3. De novo assembly of the E. uniflora transcriptome

All low quality reads with FASTQ values below 13 were removed, and 5′ and 3′ adapter sequences were trimmed using the Genome Analyzer Pipeline (Illumina) at Fasteris SA. Error correction of reads containing 'n' was performed with ALLPATHS-LG [25]. High quality and corrected reads were assembled using Velvet/Oases software [26] with multiple k-mers (21, 31, 41, 51 and 61). The total contigs obtained in each different assembly were merged to produce a combined assembly using the USEARCH algorithm [27]. After this step, we obtained unified contigs or unigenes with a minimum length of 200 nt. To compare the performance and quality of the unigenes obtained with different k-mers using Velvet/Oases, we also performed other de novo transcriptome assemblies using Trinity [28] and CLC Genome Workbench version 4.0.2 (CLCbio, Aarhus, Denmark) software. After assembly, the USEARCH algorithm was used to eliminate redundant contigs in each transcriptome and to obtain unigenes. Various parameters, including the overall number of contigs, the average length of contigs and the N50 value (the median contig length), were used to compare the different assemblies obtained.

2.4. Annotation of gene families, protein domains and functional classification

The unigenes obtained with Velvet/Oases were compared with non-redundant sequences from NCBI (http://www.ncbi.nlm.nih.gov/) and the Swiss-Prot (http://www.expasy.ch/sprot) databases using BLASTX [29] with an *E*-value cutoff of 10^{-6}. The best hits of each unigene with the highest sequence similarity were chosen, and the annotations were obtained using the Perl script developed by Sloan et al. [30]. To determine unigene abundances, high quality reads were mapped to the annotated unigenes using Bowtie software [31]. A functional category assignment for each unigene was conducted using the GOslim tool from the Blast2GO suite [32], and classification was performed according to GO terms within molecular functions, biological processes and cellular components. The identification of gene families and protein domains was performed using the InterProScan tool from the Blast2GO suite from multiple databases, including Gene3D, PANTHER, Pfam, PIR, PRINTS, ProDom, ProSITE, SMART, SUPERFAMILY and TIGERFAM.

2.5. Pathway assignment with KEGG

Pathway mapping of the unigenes using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (http://www.genome.jp/kegg/) was performed with the Blast2GO suite. The unigenes were annotated to the KEGG database to obtain their enzyme commission (EC) number. This code was further used to map the unigenes to the KEGG biochemical pathways. We focused on unigenes whose function was assigned to the terpenoid backbone biosynthesis pathway. These unigenes were manually curated by translating their nucleotide sequences into peptides after querying the longest predicted open reading frame (ORF) using the ORF Finder server (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and comparing their sequence length to other similar peptide sequences identified in other plants using BLASTP.

2.6. Identification of terpene synthases and oxidosqualene cyclases

The TPS and OSC genes from *E. uniflora* were identified using the domain/family annotation previously obtained with InterProScan. Because we are interested in the identification of full-length sequences, only those unigenes containing the Pfam families PF01397 (terpene synthase N terminal domain) and PF03936 (terpene synthase C terminal domain) were considered as TPS genes. In the case of the OSC genes, we only considered the unigenes containing the InterPro family/domains IPR018333 (squalene cyclase) and IPR008930 (terpenoid cyclases/protein prenyltransferase alpha-alpha toroid). These identified unigenes were manually curated as previously described. As several groups of homologous TPS and OSC were identified in other plants, we used these sequences as references to performed a phylogenetic analysis with the identified genes to infer, by analogy, the potential type of terpene synthetized by the genes identified in this study [33]. A phylogenetic analysis was performed separately for TPS and OSC genes after protein and nucleotide sequence alignment using a Bayesian method. For this analysis, sequence of TPS and OSC genes previously identified and characterized in 35 plants (Table S1) were used as references the, with additional TPS and OSC identified in the first annotation of the *E. grandis* genome [16]. The *E. grandis* sequences were included in the phylogenetic analysis because it is a member of Myrtaceae family. TPS and OSC genes with partial sequences, identified in the *E. uniflora* transcriptome, were also included in the analysis. All alignments were performed using MUSCLE [34] implemented in MEGA 5.1 [35] with default parameters. The Bayesian analyses were performed using BEAST1.7 software [36] with both protein and nucleotide sequence alignment. The model of protein evolution used in this analysis was the Dayhoff model, which was selected using ProtTest3 [37]. For nucleotide sequences, the substitution model was selected using Akaike's information criterion (wAIC) in the software jModelTest 2 [38], which is a GTR + I + G model. The Yule process was selected as a tree prior to Bayesian analysis, and 10,000,000 generations were performed using Markov chain Monte Carlo (MCMC) algorithms. Tracer 1.5 (http://beast.bio.ed. ac.uk/Tracer) was used to check for convergence of the Markov chains and adequate effective sample sizes (>200). The trees were visualized and edited using FIGTREE version 1.3.1 software (http://tree.bio.ed.ac.uk/software/figtree/).

2.7. Expression analysis of terpene synthases and oxidosqualene cyclases

The cDNA synthesis was performed from approximately 1 μg of total RNA. Each reaction was primed with 1 μM dT36V oligonucleotide (Invitrogen, Carlsbad, CA, USA). Before transcription, RNA and the dT26V primer oligo were mixed with RNase-free water to a total volume of 10 μL and incubated at 70 °C for 5 min followed by cooling on ice. Then, 6 μL of 5× RT buffer, 1 μL of 5 mM dNTP (Ludwig, Porto Alegre, RS, Brazil) and 1 μL MML-V RT Enzyme 200 U (Promega, Madison, WI, USA) were added for a final volume of 30 μL. The synthesis was performed at 40 °C for 60 min. All cDNA samples were diluted 100-fold with RNase-free water before use as a template in RT-PCR analysis.

Reverse transcription quantitative polymerase chain reaction (RT-qPCR) amplification was performed to validate and investigate the expression of three different classes of terpenoid genes in *E. uniflora*, the monoterpenes synthase (C10) (Eun-06523), sesquiterpenes synthase (C15) (Eun-12647) and oxidosqualene cyclase (C30) (Eun-03099) groups, across different tissues. The primers were designed based on transcriptome assembly and are presented in table S2. All RT-qPCR reactions were performed in a Bio-Rad CFX384 real-time PCR detection system (Bio-Rad, Hercules, CA, USA) for mature/young leaves and petal tissues using SYBR Green I (Invitrogen, Carlsbad, CA, USA) to detect double-stranded cDNA synthesis. Reactions were conducted in a volume of 10 μL containing 5 μL of diluted cDNA (1:100), 1X SYBR Green I, 0.025 mM dNTP, 1X PCR buffer, 3 mM $MgCl_2$, 0.25 U Platinum Taq DNA Polymerase (Invitrogen, Carlsbad, CA, USA) and 200 nM of each reverse and forward primer. Samples were analyzed in biological triplicates and technical quadruplicates in a 384-well plate, and a no-template control was also included. The PCR reactions were run as follows: an initial polymerase hot start step for 5 min at 94 °C and 40 cycles of 15 s at 94 °C, 15 s at 60 °C and 10 s at 72 °C. A melting curve analysis was programmed at the end of the PCR run over the range of 65–99 °C, and the temperature increased stepwise by 0.5 °C. The threshold and baselines were manually determined using the Bio-Rad CFX manager software. To calculate the relative expression of the TPS and OSC genes, we used the $2^{-\Delta\Delta Ct}$ method [39]. A Kruskal–Wallis statistical test was performed to compare the differences in expression among the different samples. The means were considered significantly different when $P < 0.01$. The products of RT-qPCR were also analyzed by electrophoresis on a 2% agarose gel stained with SYBR Gold (Applied Biosystems, Foster City, CA, USA) and visualized using UV light.

3. Results

3.1. Outputs of E. uniflora transcriptome de novo assembly

To obtain the *E. uniflora* transcriptome, an RNAseq library was constructed from leaves and sequenced using Illumina. A total of 1,676,576,700 bases from 16,765,767 sequence reads were obtained (Table 1). These raw data were assembled into 304,425 contigs using multiple k-mers with Velvet/Oases. Therefore, 74,231, 65,955, 59,704, 55,479, and 49,056 contigs were obtained using 21, 31, 41, 51 or 61 k-mers, respectively, in each de novo assembly (Table S3). These multiple k-mer assemblies were further merged with USEARCH into 72,742 unigenes. The mean unigene size was 1048 bp with lengths ranging from 200 to 10,204 bp, and the mean N50 was 1640 bp. The unigene length distribution showed that approximately 39% of the unigenes contained more than 1000 bp (Fig. 1). A statistical comparison of the unigenes obtained with Velvet/Oases and other transcriptome assembly tools showed that the use of different k-mers improved the mean unigene length, N50 and the number of unigenes with more than 1000 bp (Table S4).

3.2. Transcriptome annotation

For unigene annotation, sequence similarity searches were conducted against the NCBI non-redundant protein (Nr) and Swiss-Prot protein databases using the BLASTX algorithm with an e-value threshold of 10^{-5}. The results showed that 43,631 (59.98%) out of 72,742 unigenes showed significant similarities to known proteins in the Nr database, and 36,289 (49.89%) unigenes had BLAST hits in the Swiss-Prot database (Table 2). Detailed descriptions of the unigene annotations for the Nr and Swiss-Prot databases are provided in Table S5. The species distribution of the top hits against the Nr database showed that 18,443 (42.27%) of the annotated unigenes

Table 1
Summary of RNA-Seq and de novo assembly of *Eugenia uniflora* transcriptome using Velvet/Oases with multiple k-mers.

Sequence	Number	Mean size	N50 size	Total nucleotides
Read	16,765,767	100	100	1,676,576,700
Contig	304,425	702	1700	306,038,388
Unigene	72,742	1048	1640	76,285,081

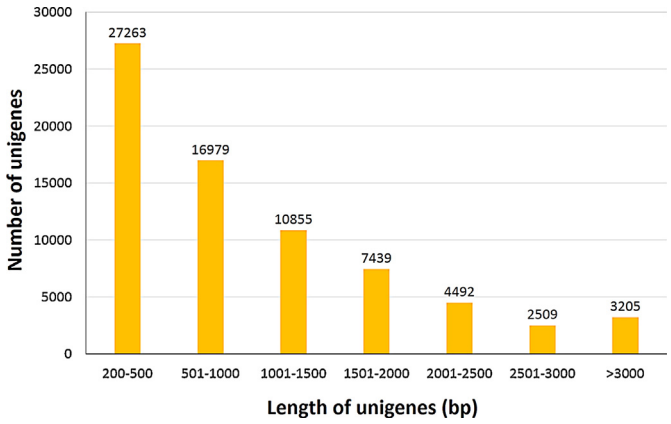

Fig. 1. Length distribution of the assembled unigenes of *E. uniflora*.

had first hits with sequences from *Vitis vinifera*, followed by *Glycine max* (13.75%), *Medicago truncatula* (8.67%) and *Arabidopsis thaliana* (8.60%) (Figure S1). On the other hand, we only found only 185 (0.42%) unigenes that matched sequences from other Myrtaceae species deposited in the Nr database. Among these sequences, 120 unigenes had first hits with sequences from *Eucalyptus grandis*, and 65 unigenes had hits with sequences from *Eucalyptus globulus*.

The top 50 annotated unigenes with the highest total reads accounted for 7.04% of the total mapped reads, and they were followed by approximately 5810 unigenes that accounted for 59.59% of the total mapped reads. The remaining 37,771 unigenes with a length lower than 1000 bp represented 33.37% of the total mapped reads. Among the most abundant and expressed genes in the *E. uniflora* transcriptome, it is possible to identify the homologs encoding cell wall-associated hydrolase, Cu/Zn-superoxide dismutase, sedoheptulose-1,7-bisphosphatase, glycine decarboxylase, Ycf68, 20S proteasome subunit alpha-1, Hop-interacting protein THI141, catalase, cytochrome P450, proteins from photosystem I and II and ribulose bisphosphate carboxylase (Table S6).

We identified several conserved protein domains/families in the *E. uniflora* unigenes using the InterPro database as a reference. The 60,952 top hits obtained were categorized into 5338 domains/families. Most categories contain from 1 to 4 unigenes that appear most frequently. The protein domains and gene families were ranked according to the number of unigenes and the fifty most abundant are provided in Table S7. The top conserved domains were categories related with different types of protein kinase domains involved in the regulation of the majority of cellular pathways. Other highly represented categories were pentatricopeptide repeat, NAD(P)-binding domain, leucine-rich repeat, zinc finger (RING/FYVE/PHD-type), tetratricopeptide-like helical, NB-ARC and WD40-repeat-containing domain.

3.3. Gene ontology classification

To functionally categorize the *E. uniflora* transcriptome, GO terms were assigned to each unigene based on their sequence matches to known protein sequences in the Nr database. Out of 72,742 unigenes, 38,314 (52.67%) were assigned at least one GO term (Table 2). Among these unigenes, 29,291 (40.26%) were in the molecular function category, 30,323 (41.68%) were in the cellular component category and 35,706 (49.08%) were in the biological process category. Based on GO annotation, *E. uniflora* unigenes were categorized into 53 functional groups. Protein binding (8832 unigenes), intracellular membrane-bounded organelle (19,724 unigenes) and primary metabolic process (15,347 unigenes) were the most abundant GO slims in each of the molecular function, cellular component and biological process categories, respectively (Fig. 2). We also noticed a high percentage of unigenes from functional groups of cellular metabolic process (11,539 unigenes), plastid (9539 unigenes), macromolecule metabolic process (9167 unigenes), biosynthetic process (8685 unigenes), transferase activity (8159 unigenes), hydrolase activity (7382 unigenes), nucleic acid binding (6589 unigenes) and mitochondrion (5197 unigenes). Additionally, we found 1302 unigenes in the functional group of secondary metabolic process.

Table 2
Summary of annotations of assembled *E. uniflora* unigenes.

Category	Number of unigenes	Percentage (%)
NCBI non-redundant annotated	43,631	59.98
Swissprot annotated	36,289	49.89
GO classified	38,314	52.67
KEGG classified	8625	11.86

3.4. Terpene backbone biosynthesis pathway mapping by KEGG

To identify active metabolic pathways in the leaves of *E. uniflora*, we mapped the unigenes to the KEGG reference pathways using Blast2GO. Out of the 72,742 assembled unigenes, 8625 (11.86%) were mapped to unique Enzyme Commission (EC) numbers (Table 2). These unigenes were assigned to 141 metabolic pathways (Table S8). Of these identified KEGG pathways, purine metabolism was the largest and contained 975 unigenes. Other pathways included starch and sucrose metabolism (829 unigenes), glycolysis/gluconeogenesis (505 unigenes), pyrimidine metabolism (430 unigenes), glyoxylate and dicarboxylate metabolism (411 unigenes), carbon fixation in photosynthetic organisms (406 unigenes), pyruvate metabolism (398 unigenes) and glycerolipid metabolism (394 unigenes).

A set of 40 unigenes codifying 16 key enzymes that control the terpene backbone biosynthesis pathway was also identified (Table S9). These unigenes are distributed in the mevalonate pathway (15 unigenes, 6 enzymes) and the MEP/DOXP pathway (14 unigenes, 7 enzymes), which are responsible for the synthesis of the terpenoid building block isopentenyl diphosphate (Fig. 3). We found only one unigene codifying the isopentenyl diphosphate delta-isomerase that catalyzes the conversion of isopentenyl diphosphate to dimethylallyl diphosphate. Additionally, we identified prenyltransferases that generate higher-order building blocks: geranyl diphosphate synthase (9 unigenes) and farsenyl diphosphate synthase (1 unigene), which are the precursors of monoterpenes, sesquiterpenes and triterpenes.

3.5. Putative terpene synthases and oxidosqualene cyclases from *E. uniflora*

Using the InterProScan search, four TPS and three OSC full-length genes were identified in the leaf transcriptome of *E. uniflora*. After comparing these unigenes to the non-redundant sequence database from NCBI using BLASTP, we found that these sequences are homologs of other plant TPS and OSC involved in monoterpene, sesquiterpenes and triterpenes biosynthesis. More specifically, we identified 1 putative monoterpene synthase (Eun-06523), 3 putative sesquiterpene synthases (Eun-11169, Eun-11883 and Eun-12647) and 3 putative oxidosqualene cyclases (Eun-04525, Eun-03099 and Eun-04273) and in the *E. uniflora* transcriptome (Table 3). In addition, we identified 7 TPS and 2 OSC genes with partial sequences (Table S10) that were also used for the phylogenetic analysis to determinate their possible function and better describe the number of genes present in the *E. uniflora* transcriptome.

A phylogenetic analysis performed with the putative TPS and OSC protein sequences from *E. uniflora* and their homologs

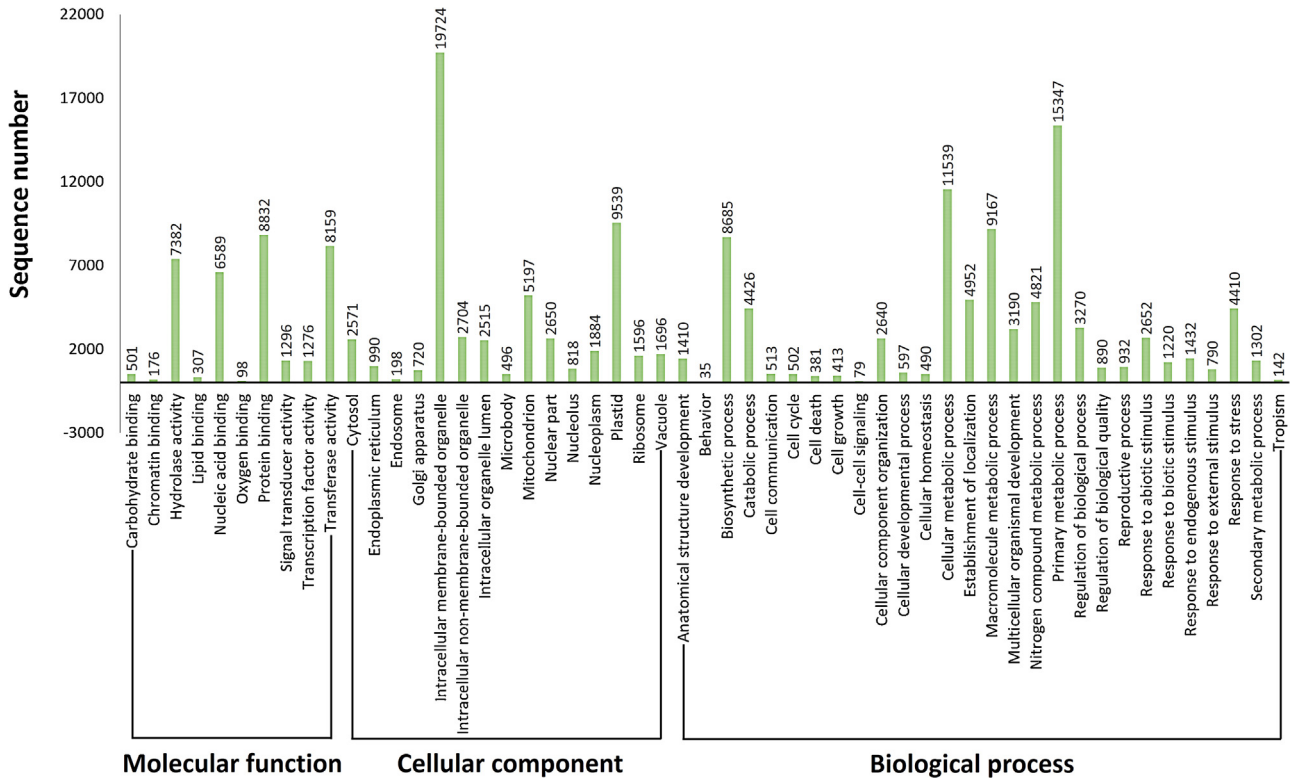

Fig. 2. Histogram of the GO classifications of annotated unigenes from the *E. uniflora* transcriptome.

identified in other plants, including *E. grandis* species, revealed a well-supported tree for each family (Fig. 4). The trees constructed with nucleotide sequences presented the same topology (data not shown). This analysis allows us to infer possible functions for the putative TPS and OSC genes that were identified in the *E. uniflora* transcriptome. The phylogenetic tree showed the formation of three main groups of homologous TPS that were grouped with relation to the type of terpene backbone produced (monoterpene synthases group, diterpene synthases group and sesquiterpene synthases group). In this manner, unigene Eun-06523 was grouped with other monoterpene synthases, and unigenes Eun-11169, Eun-11883 and Eun-12647 were grouped with sesquiterpene synthases. Two diterpenes synthases with partial sequences (Eun-45582 and Eun-24593) were also identified among the terpene synthases from the *E. uniflora* transcriptome. Similar to the TPS and OSC sequences from *E. uniflora*, the TPS and OSC genes from the first annotation of the *E. grandis* genome were also distributed in both threes. The sequences of the seven putative genes involved in the biosynthesis of terpenoids from *E. uniflora* identified in this study were submitted to the GenBank database, and the accession numbers are available in Table 3.

3.6. Putative terpene synthases and oxidosqualene cyclases expression profile by RT-qPCR

The expression of TPS and OSC genes from *E. uniflora*, including monoterpenes synthase (Eun-06523), sesquiterpenes synthase (Eun-12647) and oxidosqualene cyclase (Eun-03099), were validated and measured by RT-qPCR. The expression of these genes was analyzed in young and mature leaves and petal tissues collected from *E. uniflora* trees. We used E3 ubiquitin ligase (E3), histone H2A (H2A) and glycerol-3-phosphate dehydrogenase (GPDH) as reference genes, which we observed to be optimal normalizers for different tissues in *E. uniflora* by geNorm analysis (Figure S2) [40]. A different expression pattern was identified for the three genes (Fig. 5). We observed that the sesquiterpene synthase gene was equally expressed in young and mature leaves, with only traces of its transcripts being detected in petal tissues. A different behavior was observed for monoterpene synthase expression, where a pronounced increase in its transcripts was observed in mature leaves, and lower levels were detected in young leaves and petal samples. The oxidosqualene cyclase was the only gene to be differentially expressed across the three tissues. For this class, the highest transcript accumulation was in mature leaves, a moderate level was

Table 3
Putative terpene synthases and oxidosqualene cyclases from *E. uniflora* involved in monoterpenes, sesquiterpenes and triterpenes biosynthesis.

Enzyme type	Unigene	ORF size (aa)	Putative annotation	Reads number
Monoterpene synthase	Eun-06523	581 aa	(E)-beta-ocimene synthase [*Malus domestica*]	7475
	Eun-11169	579 aa	(−)-Germacrene D synthase [*Vitis vinifera*]	115
Sesquiterpene synthase	Eun-11883	582 aa	(E)-beta-caryophyllene synthase [*Vitis vinifera*]	125
	Eun-12647	593 aa	(−)-Germacrene D synthase [*Vitis vinifera*]	234
	Eun-04525	761 aa	Beta-amyrin synthase [*Betula platyphylla*]	994
Oxidosqualene cyclase	Eun-03099	757 aa	Cycloartenol synthase protein [*Azadirachta indica*]	651
	Eun-04273	758 aa	Beta-amyrin synthase [*Betula platyphylla*]	1235

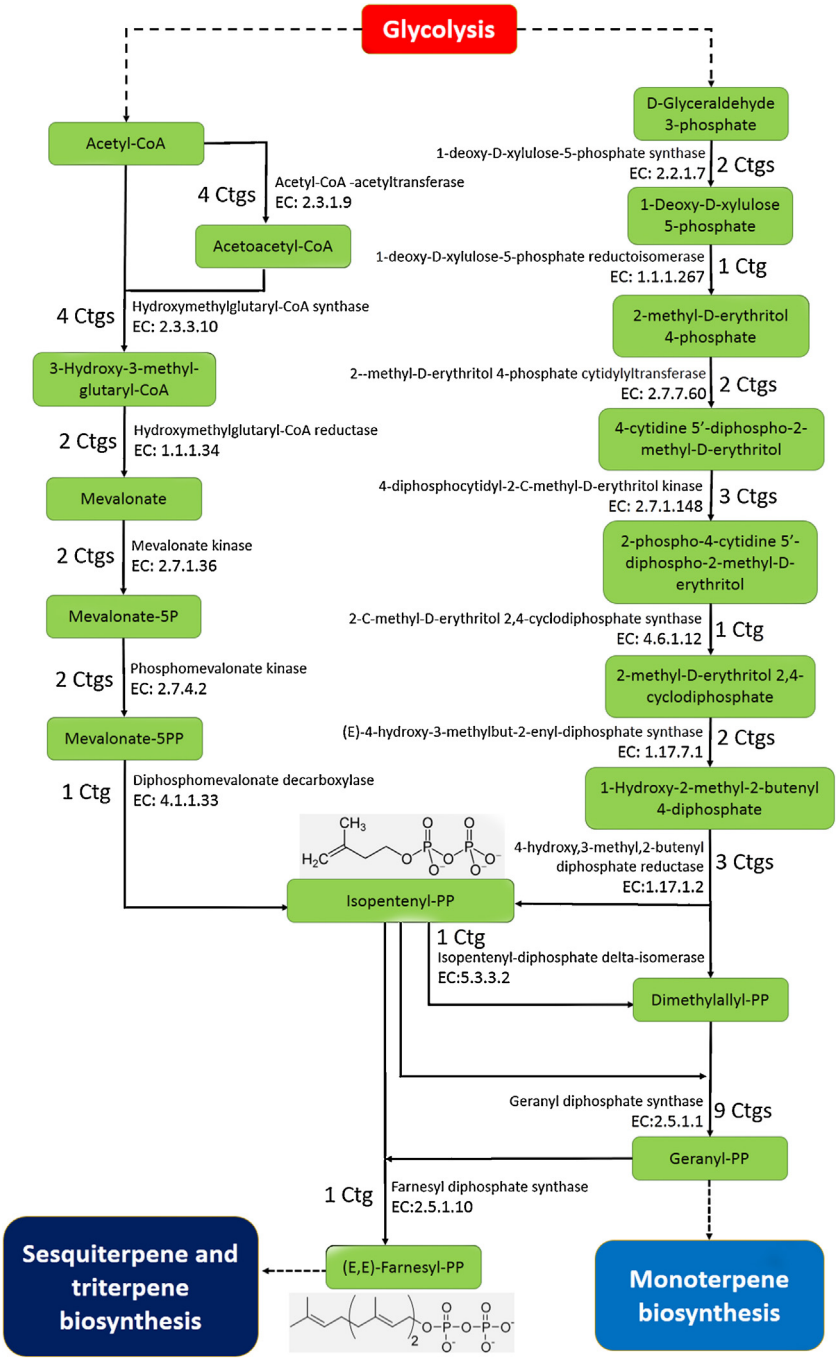

Fig. 3. Metabolic pathways showing *E. uniflora* unigenes involved in terpene backbone biosynthesis.

observed in petals, and the lowest accumulation was observed in young leaves.

To further validate the TPS and OSC transcripts, we ran the qPCR products on an agarose gel. This procedure was employed to verify the predicted amplicon size (Table S2), which was predicted after the transcript assembly of the three TPS. The expected sizes, which were 121 nt, 145 nt and 132 nt for the sesquiterpenes synthase, monoterpenes synthase and oxidosqualene cyclase, respectively, were visualized and confirmed by agarose gel analysis (Figure S3).

4. Discussion

RNA deep sequencing is a useful approach for obtaining a complete set of transcripts in a tissue from an organism at a specific developmental stage and under different physiological conditions [41]. For this reason, numerous transcriptomes from non-model plants have recently been sequenced using different next-generation sequencing (NGS) technologies in combination with multiple bioinformatics approaches [42,43].

Prior to this study, only 36 sequences from *E. uniflora* had been reported in the NCBI nucleotide database: 9 sequences of microsatellite markers [44] and 27 sequences of genes used in phylogenetic studies [45–53]. Recently, our group identified a set of 42 pre-miRNA sequences and their targets in *E. uniflora* [54]. This number is very small in comparison with the number of nucleotide sequences from *Eucalyptus* species. In this study, 16.76 million high quality reads were obtained using the HiSeq2000 platform and were assembled into 72,742 unigenes with multiple k-mers

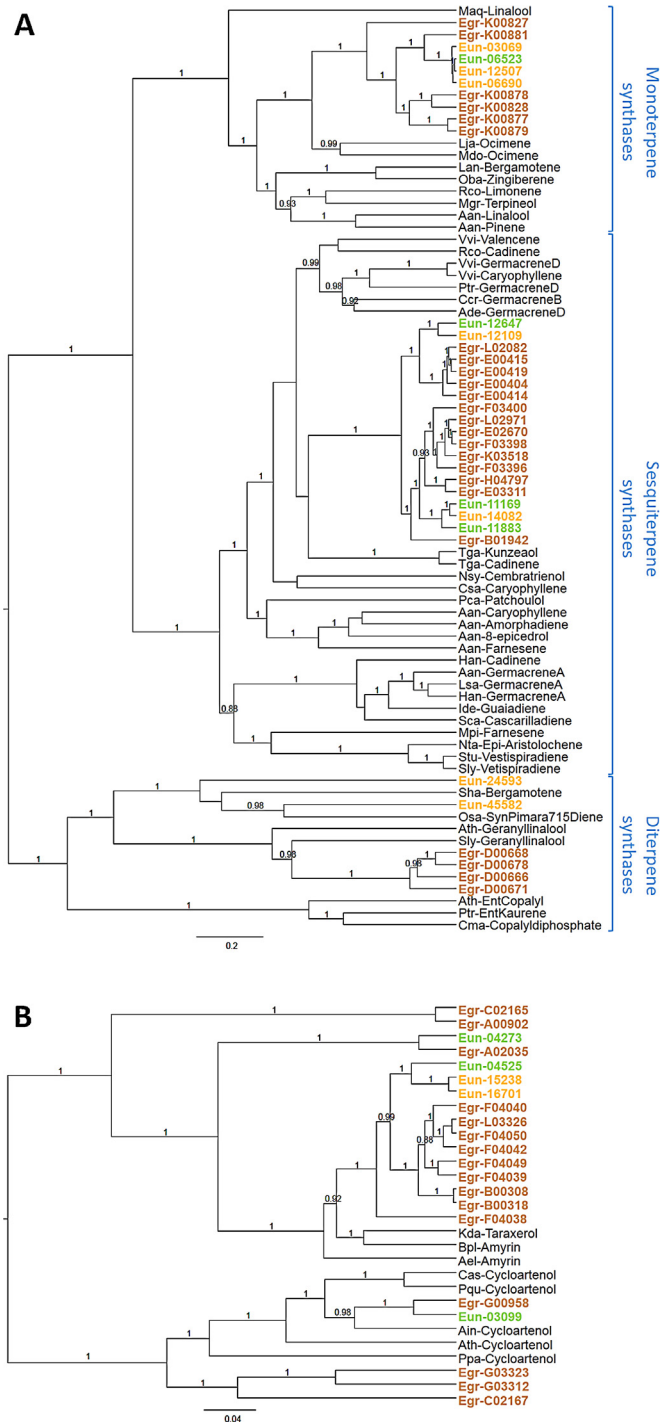

Fig. 4. Phylogenetic relationship among (A) terpene synthase and (B) oxidosqualene cyclase protein sequences reconstructed by the Bayesian method. The well-characterized TPS and OSC sequences from different plant species were used in this analysis to identify the position of *E. uniflora* and *E. grandis* among the different classes of TPS and OSC. The species used for this analysis are demonstrated in Table S1. In green: TPS and OSC full-length genes of *E. uniflora*. In orange: TPS and OSC of *E. uniflora* with partial sequences. In brown: TPS and OSC full-length genes of *E. grandis*. The posteriori probabilities are labeled above the branches.

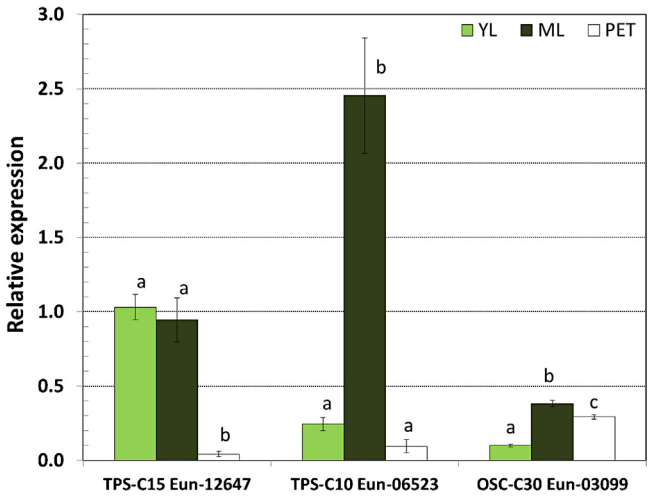

Fig. 5. Relative expression of terpene synthase and oxidosqualene cyclase genes by RT-qPCR in *E. uniflora*. Comparative analysis of sesquiterpenes synthase (TPS-C15 Eun-12647), monoterpene synthase (TPS-C10 Eun-06523) and oxidosqualene cyclase (OSC-C30 Eun-03099) transcripts across young leaves (YL), mature leaves (ML) and petal (PET) tissues. The same letters "a" or "b" indicate no significant differences among the different tissues for each TPS (Kruskal–Wallis test, $P < 0.01$).

to improve the sensitivity, especially against low expressed genes [55]. This type of approach was successfully used to de novo assemble the transcriptomes of black pepper [56] and cabbage [57]. For the species distribution of annotated unigenes, 42.27% had first hits with sequences from *V. vinifera*, which is a species with deposited and freely available genomic resources in the Nr database. Similar results were also obtained in *Paeonia suffruticosa* [58], *Corylus mandshurica* [59] and *Salvia splendens* [60]. These results demonstrate the need to generate sequences from *E. uniflora* and other Myrtaceae species and make these sequences freely available.

Functional annotation is a process of association between a group of unigenes with a network of interacting molecules in the cell to provide predicted information about the molecular functions, cellular components, biological processes and biosynthesis pathways [61]. In this study, a large number of unigenes were assigned to a wide range of GO categories, suggesting that the unigenes represent a wide diversity of transcripts from the *E. uniflora* genome. Most of the unigenes assigned with a cellular component were localized in organelles, such as the intracellular membrane-bounded organelle, plastids and mitochondrion. In the molecular function category, the annotated unigenes were mainly mapped to protein binding, transferase activity, nucleic acid binding and hydrolase activity. The biological processes are represented primarily by the metabolic processes that occur in the cell, such as primary metabolic, cellular metabolic, macromolecular metabolic and biosynthetic processes. Additionally, the annotated unigenes were used to identify the metabolic pathways present in *E. uniflora*. A graphical examination of mapped EC numbers from KEGG indicates that we have represented the majority of metabolic pathways in the transcriptome obtained in the present study. Although functional annotation provides predicted functions for these unigenes, future studies are still required for functional validation.

As *E. uniflora* is a species that produces different types of terpenes in its essential oils with important biological activities [4], we investigated in more detail the terpene backbone biosynthesis pathway. The terpene biosynthetic pathway has been well studied, and the genes that participate in this pathway have been identified in other plants [17,62]. In this study, we found all essential structural genes for the mevalonate and MEP/DOXP pathways. The assembled sequences of these unigenes were full-length cDNAs, and the predicted protein sequences were complete. Additionally, the genes of these pathways in *E. uniflora* were highly similar to those from other dicots and to those identified in another Myrtaceae [17]. A comparison among the number of identified unigenes from the two pathways in *E. uniflora* with the number of transcripts from the homologs genes in *A. thaliana* (www.arabidopsis.org) showed that probably some genes have other loci or express more

alternative transcripts. Similar results were found in the unigenes from the mevalonate and MEP/DOXP pathways in *Litsea cubeba* [63].

Though several TPS and OSC have been found in different plants with available genomic sequences, a small number of these genes have been identified in plants without genomic sequences. Using Illumina sequencing and de novo assembly, 14 TPS have been identified in *Litsea cubeba* [63] and another 17 in *Thapsia laciniata* [64]. In this study using bioinformatics approaches and phylogenetic analysis, we were able to identify full lrnght sequences corresponding to 4 putative TPS (3 sesquiterpene synthases and 1 monoterpene synthase) and 3 putative OSC in *E. uniflora*. An analysis of expression by RT-qPCR confirmed that one member of the three identified groups was expressed in different tissues. Interestingly, the putative sesquiterpene synthase Eun-12647 was more highly expressed in young leaves, and the putative monoterpene synthase Eun-06523 was more highly expressed in mature leaves, suggesting that the developmental stages of leaves might contribute to the differential abundance of terpenes in *E. uniflora* that was observed in previous studies where the leaf age was not considered [6,10,11]. Future biochemical characterization of the different TPS and OSC described here will determine specifically the type of terpene synthesized in certain conditions. The information obtained in this study will serve as a reference for genomic and genetic studies about the molecular mechanisms behind the chemical composition of the essential oils from the leaves and fruits of *E. uniflora* individuals and other Myrtaceae species.

Acknowledgments

This work was sponsored by a Productivity and Research Grant (307868/2011-7) from the National Council for Scientific and Technological Development (CNPq, Brazil). FRK was sponsored by FAPERGS/CAPES-DOCFIX (1634-2551/13-9) grant.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.plantsci.2014.10.003.

References

[1] A.G.V. Costa, D.F. Garcia-Diaz, P. Jimenez, P.I. Silva, Bioactive compounds and health benefits of exotic tropical red–black berries, J. Functional Foods 5 (2013) 539–549.
[2] F. Salgueiro, D. Felix, J.F. Caldas, M. Margis-Pinheiro, R. Margis, Even population differentiation for maternal and biparental gene markers in *Eugenia uniflora*, a widely distributed species from the Brazilian coastal Atlantic rain forest, Diversity Distribut. 10 (2004) 201–210.
[3] A.E. Consolini, O.A.N. Baldini, A.G. Amat, Pharmacological basis for the empirical use of *Eugenia uniflora* L. (Myrtaceae) as antihypertensive, J. Ethnopharmacol. 66 (1999) 33–39.
[4] T.K. Lim, Edible Medicinal And Non-Medicinal Plants: Volume 3, Fruits, Springer, Germany, 2012.
[5] A. Padovan, A. Keszei, C. Külheim, W.J. Foley, The evolution of foliar terpene diversity in Myrtaceae, Phytochem. Rev. (2013).
[6] A.T. Henriques, M.E. Sobral, A.D. Cauduro, E.E.S. Schapoval, V.L. Bassani, G. Lamaty, C. Menut, J.M. Bessière, Aromatic Plants from Brazil. II. The chemical composition of some *Eugenia* essential oils, J. Essential Oil Res. 5 (1993) 501–505.
[7] J.L. Bicas, G. Molina, A.P. Dionísio, F.F.C. Barros, R. Wagner, M.R. Maróstica, G.M. Pastore, Volatile constituents of exotic fruits from Brazil, Food Res. Int. 44 (2011) 1843–1855.
[8] F.N. Victoria, E.J. Lenardao, L. Savegnago, G. Perin, R.G. Jacob, D. Alves, W.P. da Silva, S. da Motta Ade, S. Nascente Pda, Essential oil of the leaves of *Eugenia uniflora* L.: antioxidant and antimicrobial properties, Food Chem. Toxicol. 50 (2012) 2668–2674.
[9] K.A. Rodrigues, L.V. Amorim, J.M. de Oliveira, C.N. Dias, D.F. Moraes, E.H. Andrade, J.G. Maia, S.M. Carneiro, F.A. Carvalho, *Eugenia uniflora* L. essential oil as a potential anti-*Leishmania* agent: effects on *Leishmania amazonensis* and possible mechanisms of action, Evidence-based Complementary Alternative Med.: eCAM 2013 (2013) 279726.
[10] M. Thambi, A. Tava, M. Mohanakrishnan, M. Subburaj, K.M. Pradeepkumar, P.M. Shafi, Composition and antimicrobial activities of the essential oil from *Eugenia uniflora* L. leaves growing in India, Int. J. Biomed. Sci. 4 (2013) 46–49.
[11] D.P. Costa, S.C. Santos, J.C. Seraphin, P.H. Ferri, Seasonal variability of essential oils of *Eugenia uniflora* leaves, J. Brazil. Chem. Soc. 20 (2009) 1287–1293.
[12] B. Singh, R.A. Sharma, Plant terpenes: defense responses, phylogenetic analysis, regulation and clinical applications, 3 Biotech 1 (2014) 1–23.
[13] F. Chen, D. Tholl, J. Bohlmann, E. Pichersky, The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom, Plant J. 66 (2011) 212–229.
[14] S. Aubourg, A. Lecharny, J. Bohlmann, Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana*, Mol. Genet. Genomics: MGG 267 (2002) 730–745.
[15] D.M. Martin, S. Aubourg, M.B. Schouwey, L. Daviet, M. Schalk, O. Toub, S.T. Lund, J. Bohlmann, Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays, BMC Plant Biol. 10 (2010) 226.
[16] D. Grattapaglia, R.E. Vaillancourt, M. Shepherd, B.R. Thumma, W. Foley, C. Külheim, B.M. Potts, A.A. Myburg, Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus, Tree Genetics Genomes 8 (2012) 463–508.
[17] H. Webb, R. Lanfear, J. Hamill, W.J. Foley, C. Kulheim, The yield of essential oils in *Melalecua alternifolia* (Myrtaceae) is regulated trough transcript abundance of genes in the MEP pathway, PLoS ONE 8 (2013) e60631.
[18] I. Abe, M. Rohmer, G.D. Prestwich, Enzymatic cyclization of squalene and oxidosqualene to sterols and triterpenes, Chem. Rev. 93 (1993) 2189–2206.
[19] M. Shibuya, Y. Katsube, M. Otsuka, H. Zhang, P. Tansakul, T. Xiang, Y. Ebizuka, Identification of a product specific β-amyrin synthase from *Arabidopsis thaliana*, Plant Physiol. Biochem. 47 (2009) 26–30.
[20] Z. Xue, L. Duan, D. Liu, J. Guo, S. Ge, J. Dicks, P. ÓMáille, A. Osbourn, X. Qi, Divergent evolution of oxidosqualene cyclases in plants, New Phytologist 193 (2012) 1022–1038.
[21] M. Xiao, Y. Zhang, X. Chen, E.J. Lee, C.J. Barber, R. Chakrabarty, I. Desgagne-Penix, T.M. Haslam, Y.B. Kim, E. Liu, G. MacNevin, S. Masada-Atsumi, D.W. Reed, J.M. Stout, P. Zerbe, Y. Zhang, J. Bohlmann, P.S. Covello, V. De Luca, J.E. Page, D.K. Ro, V.J. Martin, P.J. Facchini, C.W. Sensen, Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest, J. Biotechnol. 166 (2013) 122–134.
[22] S. Zenoni, A. Ferrarini, E. Giacomelli, L. Xumerle, M. Fasoli, G. Malerba, D. Bellin, M. Pezzotti, M. Delledonne, Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq, Plant Physiol. 152 (2010) 1787–1795.
[23] K.A. Wilson, A taxonomic study of the genus *Eugenia* (Myrtaceae) in Hawaii, Pacific Sci. 11 (1957) 161–180.
[24] G. Gambino, I. Perrone, A. Gribaudo, Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants, Phytochem. Anal.: PCA 19 (2008) 520–525.
[25] S. Gnerre, I. Maccallum, D. Przybylski, F.J. Ribeiro, J.N. Burton, B.J. Walker, T. Sharpe, G. Hall, T.P. Shea, S. Sykes, A.M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E.S. Lander, D.B. Jaffe, High-quality draft assemblies of mammalian genomes from massively parallel sequence data, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 1513–1518.
[26] M.H. Schulz, D.R. Zerbino, M. Vingron, E. Birney, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, Bioinformatics 28 (2012) 1086–1092.
[27] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26 (2010) 2460–2461.
[28] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. Leduc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat. Protocols 8 (2013) 1494–1512.
[29] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[30] D.B. Sloan, S.R. Keller, A.E. Berardi, B.J. Sanderson, J.F. Karpovich, D.R. Taylor, De novo transcriptome assembly and polymorphism detection in the flowering plant *Silene vulgaris* (Caryophyllaceae), Mol. Ecol. Resour. 12 (2012) 333–343.
[31] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[32] A. Conesa, S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676.
[33] P. Rabe, J.S. Dickschat, Rapid chemical characterization of bacterial terpene synthases, Angew. Chem. Int. Ed. 52 (2013) 1810–1812.
[34] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucl. Acids Res. 32 (2004) 1792–1797.
[35] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, Mol. Biol. Evol. 28 (2011) 2731–2739.
[36] A.J. Drummond, A. Rambaut, BEAST. Bayesian evolutionary analysis by sampling trees, BMC Evol. Biol. 7 (2007) 214.

[37] F. Abascal, R. Zardoya, D. Posada, ProtTest: selection of best-fit models of protein evolution, Bioinformatics 21 (2005) 2104–2105.
[38] D. Darriba, G.L. Taboada, R. Doallo, D. Posada, jModelTest 2: more models, new heuristics and parallel computing, Nat. Methods 9 (2012) 772.
[39] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method, Methods 25 (2001) 402–408.
[40] J. Vandesompele, K.D. Preter, F. Pattyn, B. Poppe, N.V. Roy, A.D. Paepe, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Genome Biol. 3 (2002) 12.
[41] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[42] S. Schliesky, U. Gowik, A.P. Weber, A. Brautigam, RNA-Seq assembly - are we there yet? Front. Plant Sci. 3 (2012) 220.
[43] M.T. Johnson, E.J. Carpenter, Z. Tian, R. Bruskiewich, J.N. Burris, C.T. Carrigan, M.W. Chase, N.D. Clarke, S. Covshoff, C.W. Depamphilis, P.P. Edger, F. Goh, S. Graham, S. Greiner, J.M. Hibberd, I. Jordon-Thaden, T.M. Kutchan, J. Leebens-Mack, M. Melkonian, N. Miles, H. Myburg, J. Patterson, J.C. Pires, P. Ralph, M. Rolf, R.F. Sage, D. Soltis, P. Soltis, D. Stevenson, C.N. Stewart Jr., B. Surek, C.J. Thomsen, J.C. Villarreal, X. Wu, Y. Zhang, M.K. Deyholos, G.K. Wong, Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes, PLoS One 7 (2012) e50226.
[44] R. Ferreira-Ramos, P.R. Laborda, M. Oliveira Santos, M.S. Mayor, M.A. Mestriner, A.P. Souza, A.L. Alzate-Marin, Genetic analysis of forest species *Eugenia uniflora* L. through of newly developed SSR markers, Conserv. Genet. 9 (2007) 1281–1285.
[45] P.G. Wilson, M.M. O'Brien, P.A. Gadek, C.J. Quinn, Myrtaceae revisited a reassessment of infrafamilial groups, Am. J. Bot. 88 (2001) 2013–2025.
[46] G. Clausing, S.S. Renner, Molecular phylogenetics of melastomataceae and memecylaceae: implications for character evolution, Am. J. Bot. 88 (2001) 486–498.
[47] M.M.v.d. Merwe, A.E.v. Wyk, A.M. Botha, Molecular phylogenetic analysis of *Eugenia* L. (Myrtaceae), with emphasis on southern African taxa, Plant Syst. Evol. 251 (2004) 21–34.
[48] E. Biffin, L.A. Craven, M.D. Crisp, P.A. Gadek, Molecular systematics of *Syzygium* and allied genera (Myrtaceae): evidence from the chloroplast genome, Taxon 55 (2006) 79–94.
[49] F. Rutschmann, T. Eriksson, K.A. Salim, E. Conti, Assessing calibration uncertainty in molecular dating: the assignment of fossils to alternative calibration points, Syst. Biol. 56 (2007) 591–608.
[50] E.J. Luca, S.A. Harris, F.F. Mazine, S.R. Belsham, E.M.N. Lughadha, A. Telford, P.E. Gasson, M.W. Chase, Suprageneric phylogenetics of Myrteae, the generically richest tribe in Myrtaceae (Myrtales), Taxon 56 (2007) 1105–1128.
[51] W.K. Soh, J. Parnell, Comparative leaf anatomy and phylogeny of *Syzygium* Gaertn, Plant Syst. Evol. 297 (2011) 1–32.
[52] J.J. Kitson, B.H. Warren, F.B. Florens, C. Baider, D. Strasberg, B.C. Emerson, Molecular characterization of trophic ecology within an island radiation of insect herbivores (Curculionidae: Entiminae: *Cratopus*), Mol. Ecol. 22 (2013) 5441–5455.
[53] F.D. Cruz, R. Margis, C.A. Mondin, A.C. Turchetto-Zolet, M. Sobral, N. Veto, M. Almerão, Phylogenetic analysis of the genus *Hexachlamys* (Myrtaceae) based on plastid and nuclear DNA sequences and their taxonomic implications, Bot. J. Linnean Soc. 172 (2013) 532–543.
[54] F. Guzman, M.P. Almerao, A.P. Korbes, G. Loss-Morais, R. Margis, Identification of microRNAs from *Eugenia uniflora* by high-throughput sequencing and bioinformatics analysis, PLoS One 7 (2012) e49811.
[55] N. Gruenheit, O. Deusch, C. Esser, M. Becker, C. Voelckel, P. Lockhart, Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants, BMC Genomics 13 (2012) 92.
[56] S.M. Gordo, D.G. Pinheiro, E.C. Moreira, S.M. Rodrigues, C. Marli, O.F.d. Poltronieri, I.T.d. Lemos, R.T. Silva, A. Ramos, H. Silva, W.A.S. Schneider, I. Sampaio Jr., S. Darnet, High-throughput sequencing of black pepper root transcriptome, BMC Plant Biol. 12 (2012) 168.
[57] H.A. Kim, C.J. Lim, S. Kim, J.K. Choe, S.H. Jo, N. Baek, S.Y. Kwon, High-throughput sequencing and de novo assembly of *Brassica oleracea* var. Capitata L. for transcriptome analysis, PLoS One 9 (2014) e92087.
[58] S. Gai, Y. Zhang, P. Mu, C. Liu, S. Liu, L. Dong, G. Zheng, Transcriptome analysis of tree peony during chilling requirement fulfillment: assembling, annotation and markers discovering, Gene 497 (2012) 256–262.
[59] H. Ma, Z. Lu, B. Liu, Q. Qiu, J. Liu, Transcriptome analyses of a Chinese hazelnut species *Corylus mandshurica*, BMC Plant Biol. 13 (2013) 152.
[60] X. Ge, H. Chen, H. Wang, A. Shi, K. Liu, De novo assembly and annotation of *Salvia splendens* transcriptome using the illumina platform, PLoS One 9 (2014) e87693.
[61] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, Data, information, knowledge and principle: back to metabolism in KEGG, Nucl. Acids Res. 42 (2014) D199–D205.
[62] D. Tholl, Terpene synthases and the regulation, diversity and biological roles of terpene metabolism, Curr. Opin. Plant Biol. 9 (2006) 297–304.
[63] X.J. Han, Y.D. Wang, Y.C. Chen, L.Y. Lin, Q.K. Wu, Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in *Litsea cubeba*, PLoS One 8 (2013) e76890.
[64] D.P. Drew, B. Dueholm, C. Weitzel, Y. Zhang, C.W. Sensen, H.T. Simonsen, Transcriptome analysis of *Thapsia laciniata* Rouy provides insights into terpenoid biosynthesis and diversity in apiaceae, Int. J. Mol. Sci. 14 (2013) 9080–9098.

Capítulo 5

Discussão e considerações finais

5. *Discussão e considerações finais*

Por serem diplóides e possuírem um genoma pequeno, as espécies da família Myrtaceae são fortes candidatas como modelos no estudo e compreensão de diferentes processos biológicos que acontecem nas plantas (Da Costa *et al*., 2008). Estas características, junto com a disponibilidade das novas tecnologias de sequenciamento e ferramentas de bioinformática, podem contribuir na obtenção de um catálogo completo de genes dessas espécies, os elementos regulatórios que controlam a expressão gênica e um marco referencial para compreender a variação genômica ao nível populacional (Feuillet *et al*., 2011). Dentro desse contexto, *E. uniflora* se apresenta como um excelente modelo no aspecto ecológico para a compreensão ao nível genético da respostas das plantas ao meio ambiente e inter-relações das mesmas com os fatores abióticos, consequência da sua adaptabilidade aos diferentes tipos de ambientes (Salgueiro *et al*., 2004). Ao mesmo tempo, a pitanga é uma espécie com um alto potencial econômico pelo consumo *in natura* de seu fruto ou por sua utilização na fabricação de sucos, sorvetes e licores, bem como pela presença de compostos fitoquímicos com diferentes tipos de propriedades farmacológicas em seus óleos essenciais (Lim, 2012).

Neste contexto, o presente estudo visou gerar recursos genômicos associados à *E. uniflora,* com a finalidade de contribuir para um melhor conhecimento da fisiologia desta espécie ao nível gênico e para disponibilizar um banco de dados dos genes e miRNAs expressos nas folhas desta espécie. O trabalho foi dividido em dois artigos: o primeiro focado na identificação dos miRNAs em *E. uniflora* envolvidos na regulação da expressão genica de alvos que são expressos nas folhas (Capítulo 3), e o segundo relacionado com o montagem *de novo* e anotação do transcriptoma da pitanga, com ênfases na identificação dos genes envolvidos na síntese dos terpenóides (Capitulo 4). Em ambos casos, para a obtenção das sequências, foram sequenciados uma biblioteca de mRNA e outra de pequenos RNAs utilizando a tecnologia Illumina, o que gerou aproximadamente 14.8 e 16.8 milhões de leituras, respetivamente.

No primeiro artigo, as análises de bioinformática dos dados gerados pelo sequenciamento dos pequenos RNAs da folha da pitanga permitiu a identificação de 204 miRNAs maduros conservados em diferentes espécies de

plantas e distribuídos em 45 famílias de miRNAs. O número e abundância dos membros desta família coincidem com o descrito em outras espécies (Zhao *et al*., 2010; Gonzalez-Ibeas *et al*., 2011; Puzey *et al*., 2012) e sugere que a ampla variação descrita é consequência da divergência funcional dos miRNAs conservados. Igualmente, o uso conjunto das duas bibliotecas sequenciadas permitiu a identificação de 25 pre-miRNAs conservados e 17 pre-miRNAs novos, que posteriormente foram validados por RT-PCR em folhas de diferentes indivíduos de pitanga. Igualmente, para conhecer a possível função dos 42 pre-miRNAs identificados, se procedeu com a identificação dos genes alvos dos miRNAs maduros mais expressos destes precursores. Os resultados obtidos foram similares ao obtidos em outras espécies de plantas (Pantaleo *et al* 2010; Colaiacovo *et al*., 2010; Gonzalez-Ibeas *et al*., 2011; Lv *et al*., 2012) e indicam que vários pre-miRNAs conservados tem como alvos fatores de transcrição envolvidos em diferentes processos biológicos de *E. uniflora*. Por outro lado, no caso dos pre-miRNAs novos ou específicos da pitanga, se observou que existem miRNAs reguladores dos genes envolvidos na resposta ao estresse abiótico e de enzimas envolvidas na síntese de metabólitos secundários. O estudo realizado representou a primeira identificação em larga escala dos miRNAs e de seus respectivos alvos em uma espécie da família Myrtaceae sem disponibilidade prévia de sequências genômicas.

No segundo artigo, a montagem *de novo* das leituras foi realizada com três ferramentas de bioinformática de uso muito comum nestes tipo de estudos: Velvet/Oases, Trinity e CLC Genomics WorkBench. A montagem feita pelo Velvet/Oases gerou um menor número de unigenes comparados àqueles obtidos com o Trinity, mas foi melhor na obtenção de unigenes de maior comprimento e com melhor valor de N50. Estes resultados são similares àqueles obtidos em outros estudos (Gordo *et al*., 2012; Kim *et al*., 2014) e confirmam que a estratégia de usar vários k-mer em uma montagem *de novo* melhora a sensibilidade na obtenção dos unigenes. A anotação funcional dos 72.742 unigenes obtidos foram associados com uma ampla variedade de categorias funcionais de GO, sugerindo que os unigenes obtidos representam uma ampla variedade de transcritos expressos em folhas da pitanga. De igual maneira, as análises feitas com o banco de dados do KEGG mostraram que

nos unigenes obtidos estão representados uma ampla variedade de enzimas da maioria das vias metabólicas que ocorrem na pitanga, em especial daquelas que fazem parte das vias do mevalonato e MEP. O estudo descreve a identificação de 4 TPS putativas com sequências completas (3 sesquiterpeno sintases e 1 monoterpeno sintase) e 3 OSC putativas com sequências completas. Estudos similares usando o sequenciamento Illumina também foram feitas em outras espécies, como no caso de *Litsea cubeba* e *Thapsia laciniata*, onde foram identificadas 14 e 17 TPS, respetivamente (Han *et al*., 2013; Drew *et al*., 2013). Análises de expressão por RT-qPCR confirmou que as TPS e OSC identificadas estão sendo expressas diferencialmente em folha jovem e madura. Estes resultados sugerem que o estádio de desenvolvimento das folhas também pode contribuir para a abundância dos terpenóides em pitanga e que não foi considerado em estudos fitoquímicos (Henriques *et al*., 1993; Thambi *et al*., 2013; Costa *et al*., 2009). Uma futura caracterização bioquímica das TPS e OSC identificadas permitirá determinar especificamente o tipo de terpenóides que são sintetizadas por elas em determinadas condições.

Além da importância destes dados na compreensão de diferentes aspectos fisiológicos e metabólicos da pitanga, os dados também serão de muita utilidade como uma referência em estudos relacionados com a estruturação das populações de pitanga e em futuros programas de melhoramento de *E. uniflora* e outras espécies de Myrtaceae.

Capítulo 6

Referências bibliográficas dos capítulos 1 e 5

REFERENCIAS BIBLIOGRÁFICAS

Abe I (2007) Enzymatic synthesis of cyclic triterpenes. Natural product reports 24:1311-1331.

Adebajo A, Oloke K, Aladesanmi A (1989) Antimicrobial activities and microbial transformation of volatile oils of *Eugenia uniflora*. Fitoterapia 60:451-455.

Aguiar RV, Cansian RL, Kubiak GB, et al. (2013) Variabilidade genética de *Eugenia uniflora L*. em remanescentes florestais em diferentes estádios sucessionais. Revista Ceres 60:226-233.

Ait-Ali T, Swain SM, Reid JB, et al. (1997) The LS locus of pea encodes the gibberellin biosynthesis enzyme ent-kaurene synthase A. The Plant Journal 11:443-454.

Amorim ACL, Lima CKF, Hovell AMC, et al. (2009) Antinociceptive and hypothermic evaluation of the leaf essential oil and isolated terpenoids from *Eugenia uniflora* L. (Brazilian Pitanga). Phytomedicine 16:923-928.

Allen E, Xie Z, Gustafson AM, Sung G-H, Spatafora JW, et al. (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. Nature Genetics 36: 1282–1290.

Ashburner M, Ball CA, Blake JA, et al. (2000) Gene Ontology: tool for the unification of biology. Nature genetics 25:25-29.

Atchison E (1947) Chromosome numbers in the Myrtaceae. American Journal of Botany:159-164.

Aubourg S, Lecharny A, Bohlmann J (2002) Genomic analysis of the terpenoid synthase (AtTPS) gene family of *Arabidopsis thaliana*. Molecular genetics and genomics 267:730-745.

Bagetti M, Facco EMP, Rodrigues DB, et al. (2009) Antioxidant capacity and composition of pitanga seeds. Ciência Rural 39:2504-2510.

Baker CH, Matsuda SP, Liu DR, et al. (1995) Molecular-cloning of the human gene encoding lanosterol synthase from a liver cDNA library. Biochemical and biophysical research communications 213:154-160.

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281-297.

Basyuni M, Oku H, Inafuku M, et al. (2006) Molecular cloning and functional expression of a multifunctional triterpene synthase cDNA from a mangrove species *Kandelia candel* (L.) Druce. Phytochemistry 67:2517-2524.

Benson DA, Cavanaugh M, Clark K, et al. (2012) GenBank. Nucleic acids research:gks1195.

Bezerra JEF, Silva-Junior JF, Lederman IE (2000) Série Frutas Nativas – Pitanga (*Eugenia uniflora* L.). FUNEP, Jaboticabal, Brasil.

Bicas JL, Molina G, Dionísio AP, Barros FFC, Wagner R, Maróstica MR, Pastore GM (2011) Volatile constituents of exotic fruits from Brazil. Food Research International 44:1843-1855.

Biffin E, Craven LA, Crisp MD, Gadek PA (2006) Molecular systematics of *Syzygium* and allied genera (Myrtaceae): evidence from the chloroplast genome. Taxon 55:79-94.

Bohlmann J, Crock J, Jetter R, et al. (1998a) Terpenoid-based defenses in conifers: cDNA cloning, characterization, and functional expression of wound-inducible (E)-α-bisabolene synthase from grand fir (*Abies grandis*). Proceedings of the National Academy of Sciences 95:6756-6761.

Bohlmann J, Meyer-Gauen G, Croteau R (1998b) Plant terpenoid synthases: molecular biology and phylogenetic analysis. Proceedings of the National Academy of Sciences 95:4126-4133.

Braga R (1985) Plantas do Nordeste, especialmente do Ceará. 4a Ed. Natal: Universitária UFNR.

Brodersen P, Sakvarelidze-Achard L, Bruun-Rasmussen M, et al. (2008) Widespread translational inhibition by plant miRNAs and siRNAs. Science 320:1185-1190.

Brooker MIH (2000) A new classification of the genus *Eucalyptus* L'Her. (Myrtaceae). Australian Systematic Botany 13:79-148.

Bruneton, J (2001) Farmacognosia Fitoquímica plantas medicinales. 2a Ed. Editorial Acribia SA, España.

Burt S (2004) Essential oils: their antibacterial properties and potential applications in foods–a review. International journal of food microbiology 94: 223–253.

Cane DE (1999) Sesquiterpene biosynthesis: cyclization mechanisms. Comprehensive natural products chemistry 2:155-200.

Calisto BM, Perez-Gil J, Bergua M, et al. (2007) Biosynthesis of isoprenoids in plants: Structure of the 2C-methyl-d-erithrytol 2, 4-cyclodiphosphate synthase from Arabidopsis thaliana. Comparison with the bacterial enzymes. Protein Science 16:2082-2088.

Chevreux B, Pfisterer T, Drescher B, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome research 14:1147-1159.

Clarke K, Yang Y, Marsh R, et al. (2013) Comparative analysis of de novo transcriptome assembly. Science China Life Sciences 56:156-162.

Clausing G, Renner SS (2001) Molecular Phylogenetics Of Melastomataceae And Memecylaceae: Implications For Character Evolution. American Journal of Botany 88:486–498.

Cloonan N, Forrest AR, Kolle G, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature methods 5:613-619.

Consolini AE, Sarubbio MG (2002) Pharmacological effects of *Eugenia uniflora* (Myrtaceae) aqueous crude extract on rat's heart. Journal of Ethnopharmacology 81:57-63.

Consolini AE, Baldini OAN, Amat AG (1999) Pharmacological basis for the empirical use of *Eugenia uniflora* L. (Myrtaceae) as antihypertensive. Journal of Ethnopharmacology 66: 33–39.

Consortium UniProt (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic acids research 41:D43-D47.

Colaiacovo M, Subacchi A, Bagnaresi P, Lamontanara A, Cattivelli L, et al. (2010) A computational-based update on microRNAs and their targets in barley (*Hordeum vulgare* L.). BMC Genomics 11: 595.

Corey E, Matsuda S, Bartel B (1993) Isolation of an *Arabidopsis thaliana* gene encoding cycloartenol synthase by functional expression in a yeast mutant lacking lanosterol synthase by the use of a chromatographic screen. Proceedings of the National Academy of Sciences 90:11628-11632.

Costa DP, Santos SC, Seraphin JC, Ferri PH (2009) Seasonal Variability of Essential Oils of *Eugenia uniflora* Leaves. Journal of the Brazilian Chemical Society 20:1287-1293

Craven L, Biffin E (2010) An infrageneric classification of Syzygium (Myrtaceae). Blumea-Biodiversity, Evolution and Biogeography of Plants 55:94-99.

Croteau R, Kutchan TM, Lewis NG (2000) Natural products (secondary metabolites). Biochemistry and molecular biology of plants 24:1250-1319.

Da Cruz F, Margis R, Mondin CA, Turchetto-Zolet AC, Sobral M, Veto N, Almerão M (2013) Phylogenetic analysis of the genus *Hexachlamys* (Myrtaceae) based on plastid and nuclear DNA sequences and their taxonomic implications. Botanical Journal of the Linnean Society 172:532–543.

Cunningham FX, Lafond TP, Gantt E (2000) Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis. Journal of bacteriology 182:5841-5848.

Cuperus JT, Fahlgren N, Carrington JC (2011) Evolution and functional diversification of MIRNA genes. The Plant Cell Online 23:431-442.

Da Costa IR, Dornelas MC, Forni-Martins ER (2008) Nuclear genome size variation in fleshy-fruited Neotropical Myrtaceae. Plant Systematics and Evolution 276:209-217.

Davis EM, Croteau R (2000) Cyclization enzymes in the biosynthesis of monoterpenes, sesquiterpenes, and diterpenes. Biosynthesis 209:53-95.

De Almeida DJ, Faria MV, Da Silva PR (2012) Biologia experimental em Pitangueira: uma revisão de cinco décadas de publicações científicas. AMBIÊNCIA 8:159-175.

Degenhardt J, Köllner TG, Gershenzon J (2009) Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. Phytochemistry 70:1621-1637.

Desmond E, Gribaldo S (2009) Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. Genome biology and evolution 1:364-381.

Dhe-Paganon S, Magrath J, Abeles RH (1994) Mechanism of mevalonate pyrophosphate decarboxylase: evidence for a carbocationic transition state. Biochemistry 33:13355-13362.

Drew DP, Dueholm B, Weitzel C, Zhang Y, Sensen CW, Simonsen HT (2013) Transcriptome Analysis of *Thapsia laciniata* Rouy Provides Insights into Terpenoid Biosynthesis and Diversity in Apiaceae. International journal of molecular sciences 14:9080-9098.

Dudareva N, Martin D, Kish CM, et al. (2003) (E)-β-ocimene and myrcene synthase genes of floral scent biosynthesis in snapdragon: function and expression of three terpene synthase genes of a new terpene synthase subfamily. The Plant Cell Online 15:1227-1241.

Fang Y, Spector DL (2007) Identification of Nuclear Dicing Bodies Containing Proteins for MicroRNA Biogenesis in Living *Arabidopsis* Plants. Current Biology 17:818-823.

Ferreira-Ramos R, Accoroni KAG, Rossi A, et al. (2014) Genetic diversity assessment for *Eugenia uniflora* L., *E. pyriformis* Cambess., *E. brasiliensis* Lam. and *E. francavilleana* O. Berg neotropical tree species (Myrtaceae) with heterologous SSR markers. Genetic Resources and Crop Evolution 61:267-272.

Ferreira-Ramos R, Laborda PR, Oliveira Santos M, Mayor MS, Mestriner MA, Souza AP, Alzate-Marin AL (2008) Genetic analysis of forest species

Eugenia uniflora L. through of newly developed SSR markers. Conservation Genetics 9:1281-1285.

Feuillet C, Leach JE, Rogers J, et al. (2011) Crop genome sequencing: lessons and rationales. Trends in Plant Science 16:77-88.

Franzon RC, Castro CM, Raseira MdCB (2010) Variabilidade genética em populações de pitangueira oriundas de autopolinização e polinização livre, acessada por AFLP. Revista Brasileira de Fruticultura, Jaboticabal 32:240-250.

Gonzalez-Ibeas D, Blanca J, Donaire L, Saladié M, Mascarell-Creus A, et al. (2011) Analysis of the melon (*Cucumis melo*) small RNAome by high-throughput pyrosequencing. BMC Genomics 12: 393.

Gordo SM et al. (2012) High-throughput sequencing of black pepper root transcriptome. BMC plant biology 12:168.

Govaerts R, Sobral M, Ashton P, et al. (2008) World checklist of Myrtaceae. Royal Botanic Gardens.

Grabherr MG, Haas BJ, Yassour M, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology 29:644-652.

Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C, Potts BM, Myburg AA (2012) Progress in Myrtaceae genetics and genomics: *Eucalyptus* as the pivotal genus. Tree Genetics & Genomes 8:463-508.

Griffiths-Jones S (2004) The microRNA registry. Nucleic acids research 32:D109-D111.

Griffiths-Jones S, Grocock RJ, Van Dongen S, et al. (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research 34:D140-D144.

Han M-H, Goud S, Song L, et al. (2004) The *Arabidopsis* double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation.

Proceedings of the National Academy of Sciences of the United States of America 101:1093-1098.

Han XJ, Wang YD, Chen YC, Lin LY, Wu QK (2013) Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in *Litsea cubeba*. PloS one 8:e76890.

Harborne JB (1984) Phytochemical Methods: A Guide to Modern Techniques of Plant Analysis. 2a Ed. Chapman and Hall, London and New York.

Hayashi H, Huang P, Takada S, et al. (2004) Differential expression of three oxidosqualene cyclase mRNAs in *Glycyrrhiza glabra*. Biological and Pharmaceutical Bulletin 27:1086-1092.

Henriques A, Sobral M, Cauduro A, et al. (1993) Aromatic plants from Brazil. II. The chemical composition of some Eugenia essential oils. Journal of Essential Oil Research 5:501-505.

Hoffmann A, Farga C, Lastra J, Veghazi E (2003) Plantas medicinales de uso común en Chile. 3a Ed. Ediciones Fundación Claudio Gay, Santiago, Chile.

Hsieh M-H, Chang C-Y, Hsu S-J, et al. (2008) Chloroplast localization of methylerythritol 4-phosphate pathway enzymes and regulation of mitochondrial genes in ispD and ispE albino mutants in *Arabidopsis*. Plant molecular biology 66:663-673.

Hunter S, Jones P, Mitchell A, et al. (2011) InterPro in 2011: new developments in the family and domain prediction database. Nucleic acids research:gkr948.

Inagaki YS, Etherington G, Geisler K, et al. (2011) Investigation of the potential for triterpene synthesis in rice through genome mining and metabolic engineering. New Phytologist 191:432-448.

Idury RM, Waterman MS (1995) A new algorithm for DNA sequence assembly. Journal of computational biology 2:291-306.

Iturbe-Ormaetxe I, Haralampidis K, Papadopoulou K, et al. (2003) Molecular cloning and characterization of triterpene synthases from *Medicago truncatula* and *Lotus japonicus*. Plant molecular biology 51:731-743.

Jenner H, Townsend B, Osbourn A (2005) Unravelling triterpene glycoside synthesis in plants: Phytochemistry and functional genomics join forces. Planta 220:503-506.

Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAS and their regulatory roles in plants. Annual Review of Plant Biology 57: 19–53.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research 42:D199-D205.

Kawano N, Ichinose K, Ebizuka Y (2002) Molecular cloning and functional expression of cDNAs encoding oxidosqualene cyclases from *Costus speciosus*. Biological and Pharmaceutical Bulletin 25:477-482.

Kim HA, Lim CJ, Kim S, Choe JK, Jo SH, Baek N, Kwon SY (2014) High-Throughput Sequencing and De Novo Assembly of *Brassica oleracea* var. Capitata L. for Transcriptome Analysis. PloS one 9:e92087.

Kitson JJ, Warren BH, Florens FB, Baider C, Strasberg D, Emerson BC (2013) Molecular characterization of trophic ecology within an island radiation of insect herbivores (Curculionidae: Entiminae: *Cratopus*). Molecular ecology 22:5441-5455.

Koyama T, Ogura K (1999) Isopentenyl diphosphate isomerase and prenyltransferases. Comprehensive natural product chemistry: isoprenoids including carotenoids and steroids 2:69-96.

Kumar S, Shah N, Garg V, et al. (2014) Large scale in-silico identification and characterization of simple sequence repeats (SSRs) from de novo assembled transcriptome of *Catharanthus roseus* (L.) G. Don. Plant Cell Reports 33:905-918.

Kurihara Y, Takashi Y, Watanabe Y (2006) The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. Rna 12:206-212.

Kurihara Y, Watanabe Y (2004) *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. Proceedings of the National Academy of Sciences of the United States of America 101:12753-12758.

MacMillan J, Beale MH (1999) Diterpene biosynthesis. Comprehensive natural products chemistry 2:217-243.

Margis R, Felix D, Caldas J, et al. (2002) Genetic differentiation among three neighboring Brazil-cherry (*Eugenia uniflora* L.) populations within the Brazilian Atlantic rain forest. Biodiversity & Conservation 11:149-163.

Margulies M, Egholm M, Altman WE, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Merchant SS, Prochnik SE, Vallon O, et al. (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245-250.

Metzker ML (2009) Sequencing technologies—the next generation. Nature Reviews Genetics 11:31-46.

Miller B, Oschinski C, Zimmer W (2001) First isolation of an isoprene synthase gene from poplar and successful expression of the gene in *Escherichia coli*. Planta 213:483-487.

Morlacchi P, Wilson WK, Xiong Q, et al. (2009) Product profile of PEN3: the last unexamined oxidosqualene cyclase in *Arabidopsis thaliana*. Organic letters 11:2627-2630.

Myburg AA, Grattapaglia D, Tuskan GA, et al. (2014) The genome of *Eucalyptus grandis*. Nature 510:356-362.

Myers EW (1995) Toward simplifying and accurately formulating fragment assembly. Journal of Computational Biology 2:275-290.

Navarro L, Dunoyer P, Jay F, et al. (2006) A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. Science 312:436-439.

Nogueira LR, Büttow MV, Castro C, et al. (2007) Avaliação da diversidade genética entre seleções de *Eugenia uniflora* através de análise da AFLP. Anais UERGS XVI 201:231-245.

Lange BM, Ahkami A (2013) Metabolic engineering of plant monoterpenes, sesquiterpenes and diterpenes—current status and future opportunities. Plant biotechnology journal 11:169-196.

Lee MH, Nishimoto S, Yang LL, et al. (1997) Two macrocyclic hydrolysable tannin dimers from *Eugenia uniflora*. Phytochemistry 44:1343-1349.

Lee Y, Kim M, Han J, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. The EMBO journal 23:4051-4060.

Li Z, Chen Y, Mu D, et al. (2012) Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Briefings in functional genomics 11:25-37.

Liang C, Zhang X, Zou J, et al. (2010) Identification of miRNA from *Porphyra yezoensis* by high-throughput sequencing and bioinformatics analysis. PloS one 5:e10698.

Lim TK (2012) Edible Medicinal And Non-Medicinal Plants: Volume 3, Fruits. Springer, Germany.

Liu J, Huang F, Wang X, et al. (2014) Genome-wide analysis of terpene synthases in soybean: functional characterization of GmTPS3. Gene 544:83-92.

Lluch MA, Masferrer A, Arró M, et al. (2000) Molecular cloning and expression analysis of the mevalonate kinase gene from *Arabidopsis thaliana*. Plant molecular biology 42:365-376.

Lorenzi H, Bacher L. Lacerda M, Sartori S (2006) Brazilian fruits & cultivated exotics (for consuming in natura). Instituto Plantarum de Etodos da Flora Ltda. Nova Odessa, Brasil.

Luca EJ et al. (2007) Suprageneric Phylogenetics of Myrteae, the Generically Richest Tribe in Myrtaceae (Myrtales). Taxon 56:1105-1128.

Luzia DM, Bertanha BJ, Jorge N (2010) Pitanga (*Eugenia uniflora* L.) seeds: antioxidant potential and fatty acids profile. Revista do Instituto Adolfo Luz 69: 175–180.

Lv S, Nie X, Wang L, Du X, Biradar SS, et al. (2012) Identification and Characterization of MicroRNAs from Barley (*Hordeum vulgare* L.) by High-Throughput Sequencing. International Journal of Molecular Sciences 13: 2973–2984.

Oguntimein B, Elakovich S (1991) Allelopathic activity of the essential oils of Nigerian medicinal plants. Pharmaceutical Biology 29:39-44.

Ohyama K, Suzuki M, Kikuchi J, et al. (2009) Dual biosynthetic pathways to phytosterol via cycloartenol and lanosterol in *Arabidopsis*. Proceedings of the National Academy of Sciences 106:725-730.

Oliveira AL, Lopes RB, Cabral FA, et al. (2006) Volatile compounds from pitanga fruit (*Eugenia uniflora* L.). Food chemistry 99:1-5.

Pantaleo V, Szittya G, Moxon S, Miozzi L, Moulton V, et al. (2010) Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. The Plant Journal 62: 960–976.

Parida SK, Kumar KAR, Dalal V, et al. (2006) Unigene derived microsatellite markers for the cereal genomes. Theoretical and applied genetics 112:808-817.

Paterson AH, Bowers JE, Bruggmann R, et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551-556.

Phillips DR, Rasbery JM, Bartel B, et al. (2006) Biosynthetic diversity in plant triterpene cyclization. Current opinion in plant biology 9:305-314.

Puzey JR, Karger A, Axtell M, Kramer EM (2012) Deep Annotation of *Populus trichocarpa* microRNAs from Diverse Tissue Sets. PloS One 7: e33034.

Ramesh SV, Ratnaparkhe MB, Kumawat G, et al. (2014) Plant miRNAome and antiviral resistance: a retrospective view and prospective challenges. Virus genes 48:1-14.

Reinhart BJ, Weinstein EG, Rhoades MW, et al. (2002) MicroRNAs in plants. Genes & development 16:1616-1626.

Rodríguez-Concepción M, Boronat A (2002) Elucidation of the methylerythritol phosphate pathway for isoprenoid biosynthesis in bacteria and plastids. A metabolic milestone achieved through genomics. Plant physiology 130:1079-1089.

Rodwell VW, Beach MJ, Bischoff KM, et al. (2000) 3-Hydroxy-3-methylglutaryl-CoA reductase. Methods in enzymology 324:259-280.

Romagnolo MB, Souza MCD (2006) O gênero *Eugenia* L.(Myrtaceae) na planície de alagável do Alto Rio Paraná, Estados de Mato Grosso do Sul e Paraná, Brasil. Acta Botanica Brasilica 20:529-548.

Rutschmann F, Eriksson T, Salim KA, Conti E (2007) Assessing Calibration Uncertainty in Molecular Dating: The Assignment of Fossils to Alternative Calibration Points. Systematic Biology 56:591-608.

Salgueiro F, Felix D, Caldas JF, Margis-Pinheiro M, Margis R (2004) Even population differentiation for maternal and biparental gene markers in *Eugenia uniflora*, a widely distributed species from the Brazilian coastal Atlantic rain forest. Diversity and Distributions 10:201-210.

Schwab W (2003) Metabolome diversity: too few genes, too many metabolites? Phytochemistry 62:837-849.

Schomburg I, Chang A, Ebeling C, et al. (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic acids research 32:D431-D433.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086-1092.

Shibuya M, Xiang T, Katsube Y, et al. (2007) Origin of structural diversity in natural triterpenes: direct synthesis of seco-triterpene skeletons by oxidosqualene cyclase. Journal of the American Chemical Society 129:1450-1455.

Simpson JT, Wong K, Jackman SD, et al. (2009) ABySS: a parallel assembler for short read sequence data. Genome research 19:1117-1123.

Singh B, Sharma RA (2014) Plant terpenes: defense responses, phylogenetic analysis, regulation and clinical applications. 3 Biotech 1:1-23.

Soh WK, Parnell J (2011) Comparative leaf anatomy and phylogeny of *Syzygium* Gaertn. Plant Systematics and Evolution 297:1-32.

Souza VC, Lorenzi H (2005) Botânica sistemática: guia ilustrado para identificação das famílias de Angiospermas da flora brasileira, baseado em APG II. Instituto Plantarum de Estudos da Flora.

Stefanello MEA, Pascoal AC, Salvador MJ (2011) Essential oils from neotropical Myrtaceae: chemical diversity and biological properties. Chemistry & biodiversity 8:73-94.

Sunkar R, Zhu J-K (2004) Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. The Plant Cell Online 16:2001-2019.

Sunkar R, Li Y-F, Jagadeeswaran G (2012) Functions of microRNAs in plant stress responses. Trends in Plant Science 17: 196–203.

Takahashi S, Kuzuyama T, Watanabe H, et al. (1998) A 1-deoxy-D-xylulose 5-phosphate reductoisomerase catalyzing the formation of 2-C-methyl-D-erythritol 4-phosphate in an alternative nonmevalonate pathway for terpenoid biosynthesis. Proceedings of the National Academy of Sciences 95:9879-9884.

Thambi M, Tava A, Mohanakrishnan M, Subburaj M, Pradeepkumar KM, Shafi PM (2013) Composition and antimicrobial activites of the essential oil from *Eugenia uniflora* L. leaves growing in India. International Journal of Biomedical Science 4:46-49.

Van der Merwe M, Van Wyk A, Botha A (2005) Molecular phylogenetic analysis of *Eugenia* L.(Myrtaceae), with emphasis on southern African taxa. Plant Systematics and Evolution 251:21-34.

Villar del Fresno AM (1999) Farmacognosia General. Editorial Síntesis, Madrid. Espanha.

Wang L, Wang M-B, Tu J-X, et al. (2007) Cloning and characterization of microRNAs from *Brassica napus*. FEBS letters 581:3848-3856.

Wilhelm BT, Marguerat S, Goodhead I, et al. (2010) Defining transcribed regions using RNA-seq. Nature protocols 5:255-266.

Wilson PG, O'Brien MM, Gadek PA, Quinn CJ (2001) Myrtaceae Revisited: A Reassessment Of Infrafamilial Groups. American Journal of Botany 88:2013-2025.

Wise ML, Croteau R (1999) Monoterpene biosynthesis. Comprehensive natural products chemistry 2:97-153.

Yamaguchi S, Saito T, Abe H, et al. (1996) Molecular cloning and characterization of a cDNA encoding the gibberellin biosynthetic enzyme ent-kaurene synthase B from pumpkin (*Cucurbita maxima* L.). The Plant Journal 10:203-213.

Xie Z, Allen E, Fahlgren N, et al. (2005) Expression of *Arabidopsis* MIRNA genes. Plant physiology 138:2145-2154.

Xin M, Wang Y, Yao Y, et al. (2010) Diverse set of microRNAs are responsive to powdery mildew infection and heat stress in wheat (*Triticum aestivum* L.). BMC plant biology 10:123.

Zenoni S et al. (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. Plant physiology 152:1787-1795.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 18:821-829.

Zhao C-Z, Xia H, Frazier TP, Yao Y-Y, Bi Y-P, et al. (2010) Deep sequencing identifies novel and conserved microRNAs in peanuts (*Arachis hypogaea* L.). BMC Plant Biology 10: 3.

Anexo 1

Identification of potential miRNAs and their targets in *Vriesea carinata* (Poales, Bromeliaceae)

Frank Guzman[a], Mauricio Pereira Almerão[b], Ana Paula Korbes[a], Ana Paula Christoff[a], Camila Martini Zanella[a], Fernanda Bered[a], Rogério Margis[a,b]

a PPGBM at Federal University of Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brazil

b Centre of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Population, Federal University of Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brazil

Artigo publicado na Plant Science (2013)

Plant Science 210 (2013) 214–223

Contents lists available at SciVerse ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Identification of potential miRNAs and their targets in *Vriesea carinata* (Poales, Bromeliaceae)

Frank Guzman[a], Mauricio Pereira Almerão[b], Ana Paula Korbes[a], Ana Paula Christoff[a], Camila Martini Zanella[a], Fernanda Bered[a], Rogério Margis[a,b,*]

[a] PPGBM at Federal University of Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brazil
[b] Centre of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Population, Federal University of Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brazil

ARTICLE INFO

Article history:
Received 8 September 2012
Received in revised form 24 April 2013
Accepted 23 May 2013
Available online 2 June 2013

Keywords:
Vriesea carinata
miRNAs
High-throughput sequencing
Stress response

ABSTRACT

The miRNAs play important roles in regulation of gene expression at the post-transcriptional level. A small RNA and RNA-seq of libraries were constructed to identify miRNAs in *Vriesea carinata*, a native bromeliad species from Brazilian Atlantic Rainforest. Illumina technology was used to perform high throughput sequencing and data was analyzed using bioinformatics tools. We obtained 2,191,509 mature miRNAs sequences representing 54 conserved families in plant species. Further analysis allowed the prediction of secondary structures for 19 conserved and 16 novel miRNAs. Potential targets were predicted from pre-miRNAs by sequence homology and validated using RTqPCR approach. This study provides the first identification of miRNAs and their potential targets of a bromeliad species.

1. Introduction

MicroRNAs (miRNAs) are small non-coding regulatory RNAs widely found in unicellular and multicellular organisms that act as regulators of gene expression at the post-transcriptional level on genes containing miRNA binding sites [1]. Mature miRNAs are single-stranded RNA molecules of approximately 21 nucleotides (nt) in length processed from a precursor molecule (pre-miRNA) [2]. To regulate protein-coding genes the mature miRNA binds in the mRNA target site leading to mRNA degradation or translation repression [3]. In plants, miRNAs have diverse biological functions and are involved in the regulation of optimal growth and development, as well as other physiological processes, including abiotic and biotic stress responses [4]. Several studies showed that many miRNAs are conserved across different plant families [5]. However, it was also reported species and family specific miRNAs that are expressed in low levels and probably have evolved recently [6].

Bromeliaceae family, with 3248 species distributed in 58 genera [7], is an example of a large and well described adaptive radiation of plant families in the Neotropics. The family is composed of terrestrial xerophytes and both facultative and obligatory epiphytes species which acquired, throughout their evolution, interesting and different adaptive mechanisms such as central tanks, CAM photosynthetic pathway, and bear absorptive trichomes, characteristics that allowed them to occupy a wide range of habitats [8,9]. Consequently, Bromeliaceae constitutes one of the most ecologically diverse and species-rich clades of flowering plants native to the New World [10]. In Brazil, the Atlantic Rainforest is considered one of the main centers of diversity and endemism of Bromeliaceae, showing 31 genera and 803 species of which 10 genera and 653 species are endemic [11]. *Vriesea carinata* is an epiphytic or terrestrial species distributed along the Brazilian Atlantic Rainforest. As a typical species of this biome [12], *V. carinata* is an interesting model for studying the expression of miRNAs in Bromeliaceae. The first step to study the expression of miRNAs is to identify miRNAs and its targets in different natural conditions. For this purpose, we performed a high-throughput sequencing analysis (Solexa technology) of small RNAs (sRNAs) from the endemic Brazilian Atlantic Rainforest species *V. carinata*.

2. Materials and methods

2.1. Plant material and RNA isolation

Total RNA was isolated from *V. carinata* leaves using Trizol reagent (Invitrogen, CA, USA), according to manufacturer's protocol. The RNA quality was evaluated by electrophoresis on a 1%

Abbreviations: miRNA, microRNA; sRNA, small RNA; pre-miRNA, microRNA precursor; MFEI, minimal folding energy index; nt, nucleotides; vca, *Vriesea carinata*.

* Corresponding author at: Centre of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Population, Building 43431, Federal University of Rio Grande do Sul – UFRGS, P.O. Box 15005, CEP 91501-970, Porto Alegre, RS, Brazil.
Tel.: +55 51 33087766; fax: +55 51 33087309.
E-mail addresses: rogerio.margis@ufrgs.br, rogerio.margis@gmail.com (R. Margis).

0168-9452/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.plantsci.2013.05.013

agarose gel and quantification was determined using a Nanodrop (Nanodrop Technologies, Wilmington, DE, USA).

2.2. Deep sequencing

Total RNA (>10 μg) from leaves was sent to Fasteris SA (Plan-les-Ouates, Switzerland) for processing. One small RNA (sRNA) library was constructed and sequenced the Illumina HiSeq2000 platform. Briefly, the construction of the small RNA libraries consisted of the following successive steps: acrylamide gel purification of the RNA fraction corresponding to the size range 20–30 nt, ligation of the 3p and 5p adapters to the RNA in two separate subsequent steps, each followed by acrylamide gel purification, cDNA synthesis followed by acrylamide gel purification, and a final step of PCR amplification to generate a cDNA colony template library for Illumina sequencing.

A polyadenylated transcript sequencing (mRNA-seq) was performed using the following successive steps: poly-A purification, cDNA synthesis using poly-T primer shotgun to generate inserts of 500 nt, 3p and 5p adapters ligations, pre-amplification, colony generation and sequencing. The Illumina output data correspondes to sequence tags of 100 bases.

2.3. Accession numbers

Sequencing data is available in Gene Expression Omnibus (GEO) under the series accession GSE38250 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38250). This accession contains the RNA-seq and sRNA libraries derived from *V. carinata* leaves.

2.4. Analysis of small RNA library

The overall procedure for analyzing Illumina small RNA library is shown in Fig. S1. All low quality reads (with FASTq value bellow 13) were removed and 5′ and 3′ adapter sequences were trimmed using Genome Analyzer Pipeline (Illumina) by Fasteris. The remaining low quality reads with 'n' were removed with PrinSeq script [13]. Sequences shorter than 18 nt and larger than 25 nt were excluded from further analysis. Small RNAs derived from Viridiplantae rRNAs, tRNAs, snRNAs and snoRNAs deposited at the tRNAdb [14], SILVA rRNA [15], and NONCODE v3.0 [16] databases and from Poales mtRNA and cpRNA deposited at NCBI GenBank database (http://ftp.ncbi.nlm.nih.gov) were identified by mapping with Bowtie [17]. After data cleaning (low quality reads, adapter sequences), mRNA-seq data was assembled *de novo* in contigs using CLC Genome Workbench version 4.0.2 (CLCbio, Aarhus, Denmark) with default parameters. In total, 41,171 contigs were assembled and used as reference for pre-miRNA and target sequence identification.

2.5. Identification of conserved and novel miRNAs

In order to determine plant conserved miRNAs, sRNA sequences were aligned with conserved non-redundant Viridiplantae miRNAs deposited at miRBase (Release 18, November 2011) using Bowtie. Complete alignment of the sequences was required and no mismatches allowed. To search for novel miRNAs, sRNA sequences were matched against a set of proprietary *V. carinata* mRNA transcript database using SOAP2 [18]. The SOAP2 output was filtered with an in house filter precursor tool to separate candidate sequences as miRNA precursors with an anchoring pattern of a block of aligned small RNAs with perfect matches. The candidate precursors obtained were confirmed manually using Tablet software [19] to visualize the anchoring pattern. As miRNAs precursors have a characteristic hairpin structure, the next step to select precursor candidates was the secondary structure analysis by RNAfold with RNAfold webserver and using annotation algorithm from the UEA sRNA toolkit [20].

We used the mfold web server (http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form) to identify the minimal folding free energy (MFE, ΔG kcal/mol) of each miRNA precursor. This value estimates the stability of the miRNA candidate-target duplex. Then, the adjust minimal folding free energy (AMFE) and the minimal folding free energy index (MFEI) were calculated according to the previous report [21]. AMFE means the MFE of a RNA sequence with 100 nt in length, which is equal to MFE/(length of a potential pre-miRNA) × 100. MFEI is equal to MFE/(length of a potential pre-miRNA)/(The percentage of nucleotides G and C). In addition, perfect stem-loop structures should have the sRNA sequence at one arm of the stem and a respective anti-sense sequence at the opposite arm. Finally, precursor candidate sequences were confirmed as novel by BLASTn algorithm from the miRBase (www.mirbase.org) and NCBI databases.

2.6. Prediction of miRNA targets

The mRNA contigs previously assembled were clustered using the Gene Indices Clustering Tools (http://compbio.dfci.harvard.edu/tgi/software/) [22] to reduce any sequence redundancy. The clustering output was passed to CAP3 assembler [23] for multiple alignment and consensus building. Contigs that cannot reach the threshold set and fall into any assembly should remain as a list of singletons. The prediction of target genes of the most abundant mature miRNAs from the conserved and novel pre-miRNAs was performed by psRNAtarget [24] using *V. carinata* assembled unigenes longer than 600 bp (default parameters and expectation value of 3.5). Candidate RNA sequences were then annotated by assignment of putative gene descriptions based on sequence similarity with previously identified genes. These genes were annotated with those details deposited in the protein database of NR and Swiss Prot/Uniprot protein database using BLASTx implemented in blast2GO v2.3.5 software [25]. The annotation was improved by analysis of conserved domains/families using InterProScan tool and Gene Ontology terms were determined by GOslim tool from blast2GO software. At the same time the orientation of the transcripts were obtained from BLAST annotations.

2.7. miRNAs and target confirmation by RT-qPCR

In order to validate the *in silico* predicted *V. carinata* miRNAs and some of their mRNAs targets, RT-qPCR reactions were performed as described in previous works [26,27]. RNA samples were extracted from two different tissues, leaf and ovary in six biological replicates, with the Trizol reagent (Invitrogen, CA, USA). The RNA quality was accessed by electrophoresis on a 1% agarose gel. Thereafter the cDNA were obtained for 17 miRNAs based on the stem-loop method [28]. Also the cDNA for mRNA targets validation were obtained based on the poli-T amplification by reverse transcription of an M-MLV RNA Polymerase (Invitrogen, CA, USA), accordingly with the manufacturer instructions. Primers used for Stem loop cDNA synthesis, mature miRNA expression and mRNA target amplification were described in Supplementary Tables S1–S3, respectively. The RT-qPCR amplifications were performed in a CFX 384 Real-Time PCR System (Bio Rad), using SYBR Green (Invitrogen). PCR reactions were carried out in a final volume of 10 μL, containing 5 μL of diluted cDNA (1:100) and 5 μL of reagents mix: 1X SYBR Green, 0.025 mM dNTP, 1X PCR buffer, 3 mM $MgCl_2$, 0.25 U Platinun Taq DNA Polimerase (Invitrogen) and 200 nM of each reverse and forward primer. The RT-qPCR conditions were set as follow: 94 °C for 5 min, 40 cycles of 94 °C for 15 s, 60 °C for 10 s and 25 s at 72 °C. In the end of the PCR run, a melting curve were evaluated. Samples were analyzed in four technical replicates, and a no

Table 1
Summary of data from sequencing of *V. carinata* small RNA libraries.

Type	Number of reads	Percentage (%)
Total reads[a]	15,986,233	100
18–25 nt	13,834,378	86
<18 nt	752,929	5
>25 nt	1,398,926	9

[a] Reads with high quality.

template negative control was included. RNA input normalizations for miRNA were performed with the miR011 and miR397, selected by geNorm [29] as the best combination of normalizers. For target mRNAs the combination of the genes vca-miR011-5p-2-t (histone acetyl tranferase mbd9) and vca-miR396-5p-2-t (nuclear pore complex protein nup98-nup96) were set as the best combination for RNA input normalization by geNorm. To calculate the relative expression of miRNAs and mRNA targets the $2^{-\Delta\Delta ct}$ method were used [30]. To compare pairwise differences in expression, Student's *t*-test was performed considering $p < 0.05$.

3. Results

3.1. V. carinata RNA library sequencing

To identify miRNAs, a sRNA library was constructed from leaves of *V. carinata*. After library sequencing, removal of adapter, insert, and short RNAs smaller than 18 nt long and longer than 25 nt long, a total of 15,986,233 reads were obtained (Table 1). The length distribution pattern and the number of reads between 18 and 25 nt (13,834,378) in the redundant and non-redundant sRNAs datasets are shown in Fig. 1 and Table S4, respectively. The highest abundance was found for sequences in the range of 21–24 nt, whereas the highest reads redundancy was observed in the 24 nt length. Around 15.84% of reads matched miRNAs, 7% matched noncoding sRNAs, (rRNA, tRNA, snRNA, snoRNA), 6% matched organellar sRNAs (mtRNA, cpRNA) and 70.96% matched other sRNAs (Table 2).

Because the genome of *V. carinata* is not publically available, we sequenced the mRNA transcriptome of *V. carinata* leaves for use as a reference sequence in further analysis. The pooled mRNA-seq yielded 21,424,214 reads, which were imported into the CLC Genomics Workbench and *de novo* assembled into 41,171 contigs with an average length of 695 bp.

Fig. 1. Length distribution of unique and redundant *V. carinata* small RNAs.

Table 2
Categorization of *V. carinata* noncoding and organellar small RNAs.[a]

Class of small RNA	Number of reads	Percentage (%)
miRNAs	2,191,509	15.84
rRNA	840,795	6.07
tRNA	131,927	0.95
snRNA	7108	0.05
snoRNA	734	0.01
mtRNA	344,331	2.48
cpRNA	500,623	3.61
Other sRNAs	9,817,351	70.96

[a] 18–25 nt reads considered.

3.2. Identification of conserved miRNAs in V. carinata

There were 4.677 miRNAs from 47 Magnoliophyta species deposited in the microRNA database (miRBase, Release 18.0, November 2011). To identify conserved mature miRNAs in *V. carinata*, sRNA library was matched against a set of 2.585 plant unique mature miRNAs from miRBase. We identified 1.101.505 reads represented by 182 conserved plant miRNAs, which were distributed in 54 conserved miRNAs families with an average of about 4 miRNA members per family (Fig. 2 and Table S5). The most abundant families were MIR166 (20 members), MIR156 (19 members), MIR396 (15 members), MIR169 (11 members), MIR167 (10 members) and MIR159 (7 members). Some of these families are often among most represented miRNA families in other plant species [31–35]. Thirty-one families were represented by only one member (Table S5).

Globally, the relative representation of each miRNA family in terms of number of reads was variable. For example, the most represented families were MIR167, MIR159 and MIR166 with ~39% (433.596 reads), ~29% (320.878 reads) and ~14% (154.768 reads) of total reads (Table S2). The frequencies of reads ranged from 1 (14 families) to 433.596 reads (MIR167), indicating that expression varied significantly among different miRNA families. In some cases, variation of number of reads was very impressive (1–433.596 reads, MIR167; 1–136.889 reads, MIR166 and 2–319.041 reads, MIR159). These results indicated that different members have clearly different expression levels in each miRNA family. Also, the high abundance could reflect the role of these miRNA families in fundamental biological process.

Since the genome of *V. carinata* is not publically available, the sRNA library was matched against a set of *de novo* assembled contigs from *V. carinata* mRNA-seq of leaves to identify putative pre-miRNA sequences. Candidate sequences with hairpin-like structure and mature miRNAs anchored in the 5p or 3p or in both arms were further considered. Our analysis allowed the identification of 19 pre-miRNA sequences grouped in 13 conserved miRNA families (Table 3). The pre-miRNA length ranged from 78 to 209 pb and the lower minimal folding free energy (MFE) ranged from −36.90 to −83.60 (an average −49.84) and minimal folding free energy index (MFEI) ranged from −0.66 to −1.24 (an average of −1.00). MFEI is a criterion for distinguishing miRNAs from other RNAs and previous studies have shown that it is more likely to be a potential miRNA if a sequence has a MFEI value ≤ −0.85 [21]. For conserved miRNA in *V. carinata*, we detected only one exception in which conserved pre-miRNA showed MFEI under the expected value (−0.66; vca-MIR168-2) and according to Zhang et al. [41], this value is expected for mRNA. Even though MFEI is an excellent parameter to identify pre-miRNA, some studies showed considerable deviation of the expected value (≥−0.85) for conserved pre-miRNA in other plant species [32,33]. All pre-miRNA predicted by RNAfold had regular stem-loop structures (Fig. S2).

Table 3
Characteristics of pre-miRNAs identified in *V. carinata* matching conserved miRNA families in other plant species.

Pre-miRNA	Precursor miRNAs					Mature miRNAs					
						5p more abundant		3p more abundant		Total reads	Total isomiRNAs
	Contig name	Length	MFE	AMFE	MFEI	Sequence member	Reads	Sequence member	Reads		
vca-MIR156-1	contig 118752	103	−58.70	−56.99	−1.05	UGACAGAAGAGAGUGAGCAC	9473	UGCUCACUUCUCUUUCUGUCAG	730	14,963	55
vca-MIR156-2	contig 1376	86	−47.20	−54.88	−1.24	UUGACAGAAGAUAGAGAGCAC	8388	GCUCUCUAUGCUUCUGUCAUC	359	8921	22
vca-MIR159	contig 140410	209	−83.60	−40.00	−0.91	GAGCUUCCUUGGGUCCAAAGA	3500	UUUGGAUUGAAGGGAGCUCUA	319,041	332,989	49
vca-MIR160	contig 146036	103	−54.40	−52.82	−0.99	UGCCUGCCUCCCUGAAUGCCA	291	ND	–	307	4
vca-MIR166-1	contig 132716	96	−52.60	−54.79	−1.01	GGAAUGUUGUCUGGUUCGAAA	240	UCGGACCAGGCUUCAUUCCCC	136,889	148,317	30
vca-MIR166-2	contig 145255	129	−46.31	−35.89	−0.84	GGAAUGUUGUCUGGCUCGAGG	1564	UCGGACCAGGCUUCAUUCCCC	136,889	149,916	34
vca-MIR166-3	contig 75684	89	−39.80	−44.72	−0.83	GGAAUGUUGUCCCGCCUCGACG	1842	UCGGACCAGGCUUCAUUCCCC	136,889	149,967	35
vca-MIR167-1	contig 126939	104	−56.10	−53.94	−1.22	UGAAGCUGCCAGCAUGAUCUGA	410,104	CAGAUCAUGUGGCAGUUUCAU	17	411,899	30
vca-MIR167-2	contig 85667	107	−38.70	−36.17	−0.76	UGAAGCUGCCAGCAUGAUCUGA	410,104	AGAUCAUGUUGCAGUUUCAUC	40	411,769	19
vca-MIR168-1	contig 84941	91	−54.90	−60.33	−0.96	UCGCUUGGUGCAGCUCGGGAA	80,471	UCCCCGCUUGCACCAACUGAA	357	87,336	61
vca-MIR168-2	contig 97534	90	−39.70	−44.11	−0.66	UCGCUUGGUGCAGCUCGGGAA	80,471	CCCGCUUGCACCAACUGAAU	124	86,314	23
vca-MIR172-1	contig 142180	130	−50.80	−39.08	−1.08	GUGGCAUCAUGAAGAUUCACA	484	AGAAUCUUGAUGAUGCUGCAU	1783	67,657	35
vca-MIR172-2	contig 69596	94	−46.80	−49.79	−1.17	CAGCAUCAUCCAGAUUCACAU	697	AGAAUCUUGAUGAUGCUGCAU	1783	68,248	53
vca-MIR393	contig 144392	79	−40.20	−50.89	−1.22	UCCAAAGGGAUCGCAUUGAUC	1017	UCAAUCGCGAUUUCUUUGGAAU	50,329	52,013	28
vca-MIR396	contig 103942	90	−36.90	−41.00	−1.05	UUCCACAGCUUUCUUGAACU	4329	GUUCAAUAAAGCUGUGGGAAA	1116	10,058	37
vca-MIR397	contig 20347	78	−37.70	−48.33	−0.99	UCAUUGAGUGCAGCGUUGAUG	1514	UCAGCCUUCACUCAAUCAUG	54	2423	16
vca-MIR408	contig 74327	104	−61.50	−59.13	−0.99	ACGGGGACGAGGUCGGGCAUG	54	UGCACUGCCUCUUCCCUGGCU	848	1953	12
vca-MIR528	contig 8393	83	−45.50	−54.82	−0.99	UGGAAGCGGCAUGCAGAGAG	16,088	UUUCCUGUGCUCGCCCCUUCC	413	17,025	27
vca-MIR535	contig 142206	94	−55.60	−59.15	−1.11	UGACGAUGAGAGAGCACCGC	46,806	GUGCUCUCUCUCGUUGUCACC	806	49,598	78

ND: Not detected.

3.3. Identification of novel miRNAs in V. carinata

Following the criteria in the identification of conserved pre-miRNAs, we obtained 17 potential novel miRNAs grouped in 16 miRNA families (Table 4). In addition to the hairpin structure, the detection of complementary antisense miRNA (miRNA*) in all pre-miRNA was a strong indication to consider these sequences as true candidates. Comparison between the mature sequences of candidate miRNAs and sequences deposited in miRBase suggest that these candidates are novel miRNAs not identified in others species and probably specific to Bromeliacea. These sequences (miRNA*) are rarely found by cloning because of their quick degradation in cells and its detection represented further evidence for the existence of mature miRNAs [36].

The novel pre-miRNA sequences in *V. carinata* ranged from 79 to 235 nt and the lower minimal folding free energy (MFE) ranged from −26.20 to −109.50 and minimal folding free energy index (MFEI) ranged from −0.87 to −1.48 (Table 4). All novel pre-miRNA identified were in accordance with the expected value (≥−0.85) [21]. Likewise to conserved pre-miRNA, all novel pre-miRNA predicted by RNAfold had regular stem-loop structures (Fig. S2).

3.4. IsomiRs sequences

We detected multiple mature miRNAs variants named isoforms or isomiRs in both arms (5p and 3p) of identified conserved and novel pre-miRNAs (Tables 3 and 4). Because they vary widely in abundance, we considered all isomiRs in each pre-miRNA even if they showed only one read. For example, vca-nMIR016 showed six isomiRs in 3p arm varying from 2 to 13 reads and six isomiRs in 5p arm with reads ranging from 1 to 17 (Fig. 3). The existence of these multiple variants reflects the variability in miRNA biogenesis, specifically in DCL cleavage sites or subsequent processing/degradation or later processing steps [37]. Although the biological functions of isomiRs remain to be determined, isomiRs from a single pre-miRNA could be a way of broadening the regulatory network [38,39].

3.5. Identification and classification of targets for pre-miRNAs identified

To understand the biological function of miRNAs in *V. carinata*, the putative targets sites of the miRNA candidates were identified by aligning the most abundant mature miRNAs of each conserved and novel pre-miRNA identified to a set of *V. carinata* assembled unigenes using psRNAtarget. Despite the use of 4 as cut-off in previous works to infer miRNA targets [40,41], we adopted the stricter cut-off threshold of 3.5. Based on this criterion, we found 145 potential targets in total (for 27 miRNA families), of which 79 were targets of conserved miRNAs (13 miRNA families) and 66 were targets of novel miRNAs (14 miRNA families). Only for three novel miRNAs families (vca-nMIR002, vca-nMIR003 and vca-nMIR016) targets were not detected. Detailed annotation results are given in Tables 5and S6.

All targets regulated by the identified conserved and novel miRNAs in this study were subjected to GO analysis to evaluate their potential functions. The categorization of these genes according to biological process, cellular component and molecular function are summarized in Fig. 4. Based on biological process, these genes were classified into 12 categories and the four most overrepresented GO terms were response to metabolic, cellular, developmental and multicellular processes, suggesting that *V. carinata* miRNAs are involved in a broad range of physiological functions. Interestingly, we found 8 and 6 candidate genes in the category of response to abiotic stimulus in conserved and novel miRNAs, respectively (Tables 5 and S6). Categories based on molecular function revealed

Table 4
Characteristics of novel pre-miRNAs identified in *V. carinata*.

pre-miRNA	Precursor miRNAs					Mature miRNAs					
	Contig name	Length	MFE	AMFE	MFEI	5p more abundant		3p more abundant		Total reads	Total isomiRNAs
						Sequence member	Reads	Sequence member	Reads		
vca-MIR001	contig 104145	178	−60.60	−34.04	−0.87	AATAATAATTCGTTGGGCCCAAAC	12	TGAACCCAACGAGTTATTATTGCT	48	120	24
vca-MIR002	contig 104145	180	−72.80	−45.50	−1.23	TAATAACTCGTTGGGTTCAAACCA	20	CGTTTCGGCCCAACGAATT	1	42	9
vca-MIR003	contig 109307	90	−26.20	−29.11	−1.09	AAAATAATGATGTGTCGTTGCATGCG	78	ATATAGCATTATTATGGTTGCATGA	4	159	17
vca-MIR004	contig 123319	88	−55.80	−63.41	−1.40	TTGTGCTTTTAGAAGCATTCCGGC	18	GGATCGGAATGCTTCTAATAGCAC	1	29	8
vca-MIR005	contig 124561	222	−109.50	−49.32	−1.23	TTCCGGTCTGGCTTACGACAT	158	TTGACAGCTTTCAAACGGTTT	2444	3035	58
vca-MIR006	contig 125668	145	−58.20	−40.14	−1.14	TCCTGGATAATACAATAATTGCGC	11	AATTATTGTATCATTTAGGGCGGA	9	31	9
vca-MIR007	contig 77681	105	−32.60	−31.05	−1.25	AGGACAGATCAGATTCATTGAATC	4	ATTCAATGAATCTATCTGTCAAT	8	23	8
vca-MIR008	contig 79286	90	−37.00	−40.66	−1.48	AGACAGATCAGGTACATTGAATCT	144	AATGTATCTATCTGTCAATAGTAA	7	174	21
vca-MIR009	contig 147042	230	−108.00	−46.96	−1.19	AGTAAAGATGGGATGGATGGCAGA	1112	AGATTCACTAGGTGGTCCATCGCT	25	1862	76
vca-MIR010	contig 127291	102	−48.10	−47.16	−1.12	TGCGGTTACGATTATAGTGGT	1027	ACTATTTCACCAACAACCCACTAA	122	5183	95
vca-MIR011	contig 72539	109	−54.30	−49.82	−1.23	TTCAGAGCATCTCGTTCAGTGC	717	TAGGTGCTCCCGAATCATCTAGG	23	1199	37
vca-MIR012	contig 17706	141	−84.60	−60.00	−1.02	TCCGCCACGTCATCTGTGTAC	99	TCACGGATGACGTGGCGGACC	2399	2992	35
vca-MIR013	contig 68461	86	−50.70	−58.95	−1.24	TTCGCAGGAGAGATGATGCCG	457	GCATCGTCTCTCCGCGCACGG	16	562	19
vca-MIR014	contig 83228	235	−82.80	−35.23	−0.88	GTTGTCCTATGTTCCGCAGA	9	TGCCAGGAGTGTAGCATAACGC	506	527	10
vca-MIR015	contig 91889	188	−75.00	−39.89	−1.34	TCCACAGGCTTTCTTGAACTA	310,363	TTCAACAAAGTTCGTCGGAAA	4962	335,732	56
vca-MIR016	contig 70741	79	−48.00	−60.76	−1.02	TCAGGAGAGATGACACCCA	169	GGTGTCACCTCTCCTGGAC	55	306	14

that the target genes were related to eight functions and the four most frequent terms were hydrolase activity, protein binding, small molecule binding and nucleic acid binding. In the category of cellular component, the analysis revealed that the target genes are expressed on a high percentage in intracellular membrane-bounded organelle and plastid.

3.6. Biological confirmation of identified miRNAs and targets

The RT-qPCR method was used to validate the expression of the most abundant mature sequences from seventeen miRNAs (vca-miR156-1, vca-miR156-2, vca-miR159, vca-miR166, vca-miR168, vca-miR393, vca-miR396, vca-miR397, vca-miR528, vca-miR535, vca-miR009, vca-miR010, vca-miR011, vca-miR012, vca-miR013, vca-miR014 and vca-miR015) and some of their targets. We confirmed that these miRNAs were expressed in leaf and ovary tissues (results not shown). For the miRNAs vca-miR166 and vca-miR393 we found that their expressions were inversely correlated with the expression of some of its targets (Fig. S3).

4. Discussion

Several miRNAs have been identified *via* computational or experimental approaches in different plant families, but there is no sequence or functional information available about miRNAs in any Bromeliaceae species. *V. carinata* is a typical representative of the Bromeliaceae in the Brazilian Atlantic Rainforest biome and therefore emerges as interesting model for this botanical family. We used Illumina technology for deep sequencing of sRNA library to identify miRNAs in this species and results are discussed.

Size profile is an important feature to distinguish miRNA from other sRNAs and most of the mature miRNAs are of 21–25 nt. The length distribution pattern indicated that the majority of the sRNA was 24 nt, accounting for ~51% of the total reads, followed by 21 nt (~19%), 22 nt (~9%) and 23 nt (~6%). This distribution pattern is highly consistent with previous studies in other plants [42,43] and differences in distribution pattern (*e.g.*, for some species 21 nt was the most abundant) may reflect the composition of the sRNAs of a given plant species according to tissue (or cell) and physiological conditions [44]. Another explanation is that molecules of 24 nt are the typical size of Dicer-digestion product and are often the most abundant endogenous plant sRNAs [45,46]. sRNAs of 24 nt are known to be involved in heterochromatin modification, especially in a genome with a high content of repetitive sequences [47,48]. The high percentage of 24 nt sRNAs found in *V. carinata* could reflect the complexity of the genome of bromeliads [49].

Computational methods have been successfully used to predict hundreds of miRNAs in a wide variety of plant species. Illumina sequencing is a powerful tool to estimate expression profiles of miRNA. This technology provides the resources to know the abundance of various miRNA families and even distinguish between different members of a given family. In our case, we found significant differences among the number and abundance of the members identified in each family, similarity observed in other studies [50,51], suggesting that this wide variation is due to a functional divergence in the conserved miRNA families. In this study, we identified 182 conserved plant miRNAs distributed in 54 miRNAs families, very similar to found for barley, using the same approach (126 miRNA distributed in 58 miRNA families) [52].

Although conserved miRNAs have been identified by sequencing and by comparison again miRNAs identified in other species, most of plant species-specific miRNAs remain unidentified due to the levels of expression of these are very low producing a small number of reads sequenced in comparison to the conserved miRNAs [53]. For this reason, we used a new approach to identify novel miRNAs

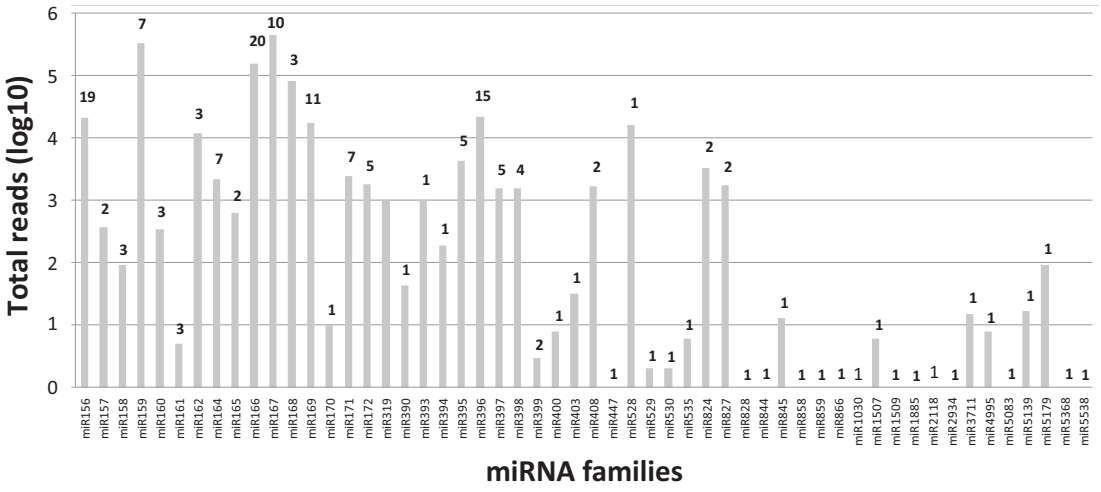

Fig. 2. Number of identified miRNAs in each conserved miRNA family in plants. The values above the bars indicate the number of members identified fot each conserved miRNA family.

in species without availability of genomic resources using simultaneously sequences from sRNA and RNAseq libraries. Using this methodology, we identified 16 potential miRNAs candidates specific for *V. carinata*. In all cases, complementary antisense miRNA (miRNA*) was identified providing more evidence for their existence as novel miRNAs [54,55].

To understand the function of the identified miRNAs, their putative targets were predicted using a bioinformatics approach.

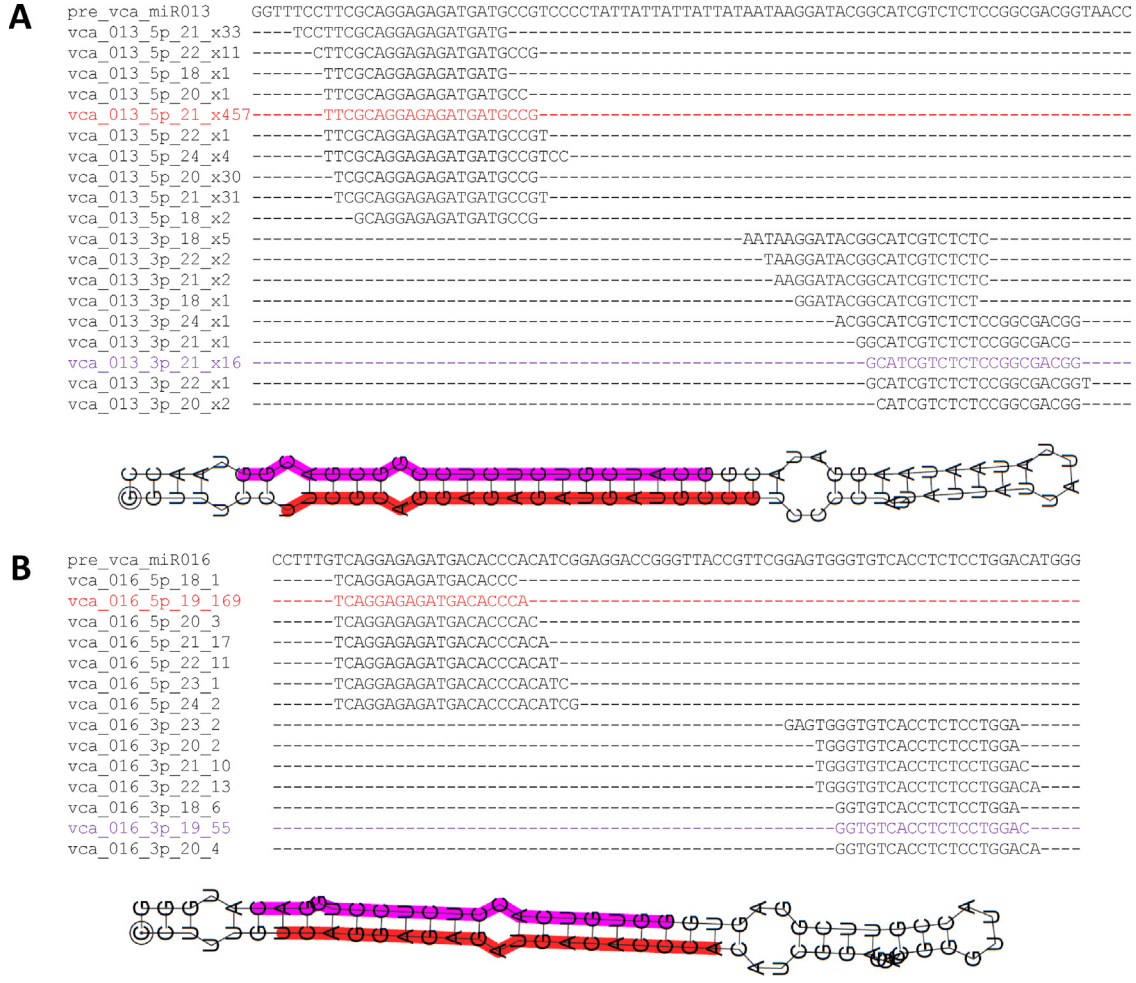

Fig. 3. Predicted secondary structures of conserved and novel miRNAs of *V. carinata*. Secondary structures of vca-nMIR013 and vca-nMIR016 precursors, their locations and the expression of small RNAs mapped onto these precursors. The sequences corresponding to the most abundant mature miRNAs in the 5p and 3p arms are labeled in red and purple, respectively. Values on the left side of the miRNA sequence represent read counts in the leaf library.

Table 5
Predicted putative targets of novel miRNAs in *V. carinata*.

Mature miRNA	Contig	Score[b]	Inhibition	Putative function
vca-miR001-3p	contig 98870	2	Cleavage	Rna binding protein
vca-miR004-5p	contig 69835_rc	2.5	Cleavage	Kinesin family member 6
vca-miR005-3p	**contig 119481**[a]	**2**	**Cleavage**	**Tetratricopeptide repeat-containing protein**
	contig 75544[a]	2.5	Cleavage	Two-component response regulator arr3
	contig 66635_rc[a]	**3**	**Cleavage**	**Heat shock 70 kda mitocondrial**
	contig 67788[a]	3	Cleavage	Pyruvate kinase isozyme chloroplastic
	contig 67948_rc	3.5	Cleavage	E3 ubiquitin ligase big brother
	contig 69145	3.5	Cleavage	Fertility restorer
	contig 95670	3.5	Cleavage	Probable lrr receptor-like serine threonine-protein kinase
	contig 143421	3.5	Cleavage	R3h domain-containing protein
	contig 70604_rc	3.5	Cleavage	Reticulon-like protein b17
	contig 67529	3.5	Cleavage	Tho complex subunit 1
vca-miR006-5p	contig 139947_rc	3	Cleavage	Lysosomal alpha-mannosidase
	contig 75839	3	Translation	Probable beta- -galactosyltransferase 2
	contig 140146	3.5	Translation	Chloroplastic group iia intron splicing facilitator chloroplastic
	contig 81980_rc	3.5	Cleavage	Omega-amidase nit2
vca-miR007-3p	contig 140165_rc	0.5	Cleavage	Calmodulin-like protein
	contig 72901_rc	3	Cleavage	Chlorophyllase- chloroplastic
	contig 68350	3	Cleavage	Soluble starch synthase chloroplastic amyloplastic
	contig 140661_rc	**3.5**	**Translation**	**Tubulin-specific chaperone d**
vca-miR008-5p	contig 69155_rc	2	Translation	Probable protein phosphatase 2c 18
	contig 140200_rc	**3**	**Cleavage**	**Abc transporter f family member 3**
	contig 71460_rc	3	Translation	Membrane protein
	contig 79377_rc	3.5	Cleavage	1-phosphatidylinositol 3-phosphate 5-kinase fab1
	contig 67906_rc	3.5	Cleavage	Aminopeptidase c
vca-miR009-5p	contig 140774_rc[a]	3	Cleavage	Mitochondrial substrate carrier family protein
	contig 118843_rc	3.5	Cleavage	Cellulose synthase
vca-miR010-5p	contig 71336[a]	1	Cleavage	Abc transporter i family member 1
vca-miR011-5p	contig 89802[a]	3	Cleavage	Dna binding protein
	contig 140009[a]	3	Cleavage	Methyl- -binding domain-containing protein 9
	contig 67880_rc	3.5	Cleavage	Phenylalanyl-trna synthetase beta subunit
	contig 122123	3.5	Cleavage	Serine threonine-protein kinase nek1
vca-miR012-3p	contig 71550_rc[a]	3	Translation	Ribonucleoside-diphosphate reductase small chain
	contig 69862[a]	3	Cleavage	Wrky transcription factor 11
	contig 34475_rc	3.5	Cleavage	Oligopeptidase b
	contig 140456_rc[a]	**3.5**	**Translation**	**Protein resurrection 1**
vca-miR013-5p	contig 139979[a]	2.5	Cleavage	Calcium-transporting atpase plasma membrane-type
	contig 66522[a]	3	Cleavage	Glucan -beta-glucosidase
	contig 66366[a]	3	Cleavage	Glutamyl-trna amidotransferase subunit a
	contig 68601[a]	3	Cleavage	Probable protein phosphatase 2c 38
	contig 118464_rc	**3.5**	**Cleavage**	**Casein kinase ii subunit beta**
	contig 705	3.5	Cleavage	Retrotransposon ty3-gypsy subclass
	contig 66929_rc	3.5	Translation	Transcription factor gte4
	contig 141231_rc	3.5	Translation	U3 small nucleolar ribonucleoprotein protein imp4
vca-miR014-3p	contig 83228_rc[a]	0	Cleavage	Nucleoid dna-binding protein cnd41
	contig 66712[a]	3	Cleavage	Uncharacterized protein ycf36
	contig 140254_rc	3.5	Cleavage	Pumilio homolog 1
vca-miR015-5p	contig 140456_rc[a]	3	Translation	Protein resurrection1
	contig 72376_rc	3.5	Translation	Ethylene-responsive transcription factor 1
	contig 140188_rc	3.5	Cleavage	Protein root hair defective 3
	contig 66270_rc	3.5	Cleavage	Ubiquitin carboxyl-terminal hydrolase 7

In bold: putative targets with response to abiotic stimulus.
[a] Targets evaluated in RTqPCR experiments.
[b] psRNATarget value.

As expected, many putative targets of conserved miRNAs were transcriptions factors, similar reported in other studies [56–58]. Among the most important transcription factors identified, we found squamosa promoter binding protein (SBP)-like (SPL) genes, which are targets of MIR156 and MIR535 families, affecting diverse developmental processes such as leaf development, shoot maturation, phase change and flowering in plants [59,60]. We also identified the auxin response factor (ARF), a plant-specific family of DNA binding proteins involved in hormone signal transduction that are targets of MIR160 and MIR167 families [61,62]. Other important genes identified and targeted by MIR156 and MIR396 were pentatricopeptide repeat genes (PPR) which show high importance in the regulation of gene expression. This gene family is implicated in post-transcriptional processes such as splicing, editing, processing and translation specifically in organelles such as mitochondria and chloroplasts [63].

In the case of the novel miRNAs, we also identified other transcription factors as targets. WRKY is the putative target of vca-nMIR012 and belongs to a large family of transcription factors that regulate various physiological processes, including pathogen defense, senescence, trichome development and signal transduction [64]. The ethylene-responsive transcription factor 1 (ESR1) was targeted by vca-nMIR015 and is involved in regulating gene expression patterns in meristems, modulating organ development and promoting shoot regeneration through the cytokinin signaling pathway [65,66]. Other important putative target identified for vca-nMIR013 was the transcription factor GTE4. It is involved in the activation and maintenance of cell division in the meristems and by this controls cell numbers in differentiated organs [67].

Based in the GO analysis, we found 14 candidate genes with response to abiotic stimulus. These results indicate that the putative targets from identified miRNAs are involved in a large number

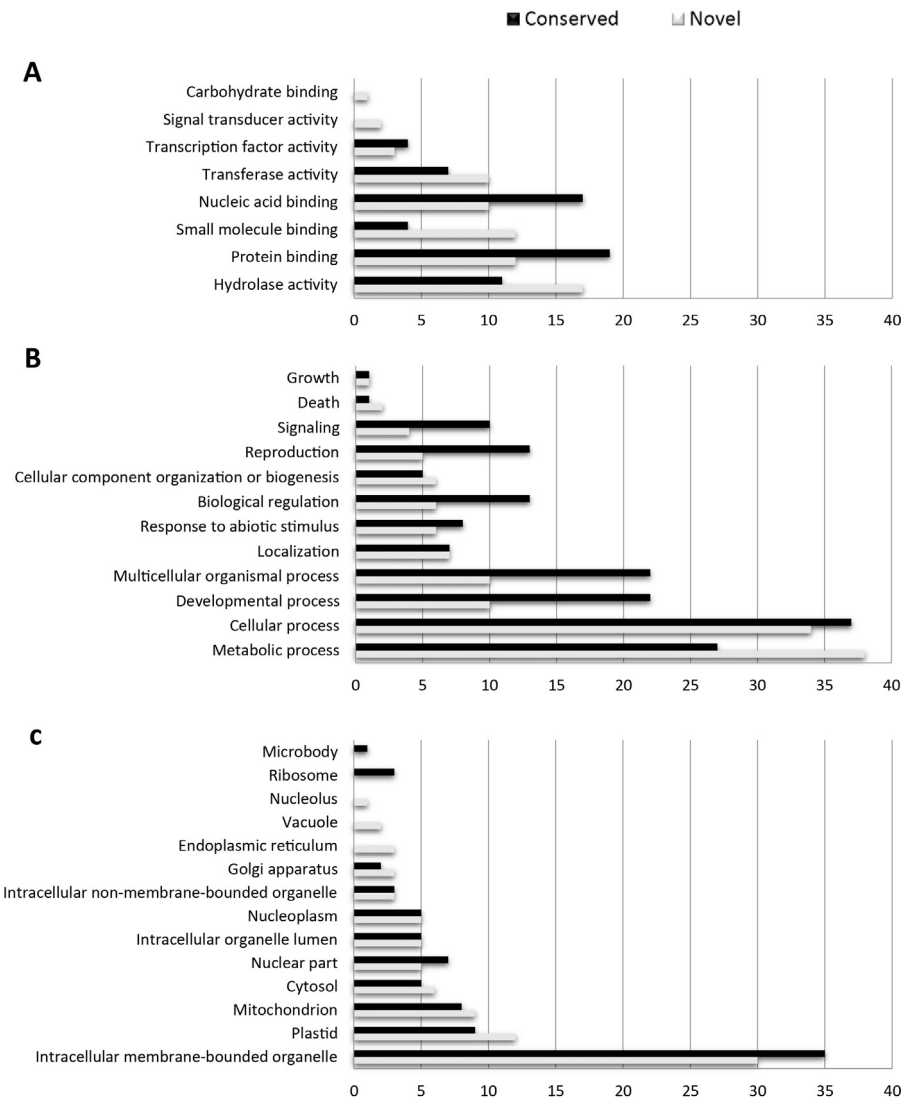

Fig. 4. Gene categories and the distribution of target genes of the most abundant mature miRNAs in the conserved and novel pre-miRNA identified in *V. carinata*. (A) Molecular function, (B) biological process and (C) cellular component.

of primary physiological and metabolic processes and are likely involved in the adaptation of *V. carinata* to different types of stress and environmental conditions observed *in natura*. Similar results in other plants confirmed the miRNA molecular regulation in different plant abiotic stresses [68]. Future experimental validation will determine how many of these predicted targets are genuinely targeted by miRNAs in *V. carinata* in specific environmental and physiological conditions.

This work provides a comprehensive investigation of the miRNA population in leaves of the bromeliad species *V. carinata*. The results support the currently idea that miRNAs are conserved in plants and play an important role in several physiological processes as shown by bioinformatics analysis of miRNA putative target genes. Although the exact function of these putative target genes remains to be confirmed, the present study provides novel insights for the comprehension of the molecular processes involved in the conserved miRNA function.

Acknowledgements

This work was sponsored by PDJ (509828/2010-8) and Productivity and Research Grant (307868/2011-7) from the National Council for Scientific and Technological Development (CNPq, Brazil).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.plantsci.2013.05.013.

References

[1] A.C. Mallory, H. Vaucheret, Functions of microRNAs and related small RNAs in plants, Nat. Genet. 38 (Suppl.) (2006) S31–S36.
[2] A.M. Denli, B.B.J. Tops, R.H.A. Plasterk, R.F. Ketting, G.J. Hannon, Processing of primary microRNAs by the Microprocessor complex, Nature 432 (2004) 231–235.
[3] M.W. Jones-Rhoades, D.P. Bartel, B. Bartel, MicroRNAS and their regulatory roles in plants, Annu. Rev. Plant Biol. 57 (2006) 19–53.
[4] R. Sunkar, Y.-F. Li, G. Jagadeeswaran, Functions of microRNAs in plant stress responses, Trends Plant Sci. 17 (2012) 196–203.
[5] T. Dezulian, J. Palatnik, D. Huson, D. Weigel, Conservation and divergence of microRNA families in plants, Genome Biol. 6 (2005) P13.
[6] E. Allen, Z. Xie, A.M. Gustafson, G.-H. Sung, J.W. Spatafora, J.C. Carrington, Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*, Nat. Genet. 36 (2004) 1282–1290.

[7] H.E. Luther, An alphabetical list of bromeliad binomials, in: Sarasota Bromeliad Society and Marie Selby Botanical Gardens, 2010.
[8] T.A. Ranker, D.E. Soltis, P.S. Soltis, A.J. Gilmartin, Subfamilial relationships of the Bromeliaceae: evidence from chloroplast DNA restriction site variation, Syst. Bot. 15 (1990) 425–434.
[9] D.H. Benzing, Bromeliaceae, in: Profile of an Adaptive Radiation, Cambridge University Press, Cambridge, 2000.
[10] T.J. Givnish, M.H.J. Barfuss, B. Van Ee, R. Riina, K. Schulte, R. Horres, P.A. Gonsiska, R.S. Jabaily, D.M. Crayn, J.A.C. Smith, K. Winter, G.K. Brown, T.M. Evans, B.K. Holst, H. Luther, W. Till, G. Zizka, P.E. Berry, K.J. Sytsma, Phylogeny, adaptive radiation, and historical biogeography in Bromeliaceae: insights from an eight-locus plastid phylogeny, Am. J. Bot. 98 (2011) 872–895.
[11] S. Porembski, G. Martinelli, R. Ohlemuller, W. Barthlott, Diversity and ecology of saxicolous vegetation mats on inselbergs in the Brazilian Atlantic rainforest, Divers. Distrib. 4 (1998) 107–119.
[12] L.B. Smith, R.J. Downs, Flora Neotropica: Monography 14, Part 2 – Tillandsioideae, The New York Botanical Garden Hafner Press, New York, 1977.
[13] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics (Oxford, Engl.) 27 (2011) 863–864.
[14] F. Jühling, M. Mörl, R.K. Hartmann, M. Sprinzl, P.F. Stadler, J. Pütz, tRNAdb 2009: compilation of tRNA sequences and tRNA genes, Nucleic Acids Res. 37 (2009) D159–D162.
[15] E. Pruesse, C. Quast, K. Knittel, B.M. Fuchs, W. Ludwig, J. Peplies, F.O. Glöckner, SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB, Nucleic Acids Res. 35 (2007) 7188–7196.
[16] S. He, C. Liu, G. Skogerbø, H. Zhao, J. Wang, T. Liu, B. Bai, Y. Zhao, R. Chen, NONCODE v2.0: decoding the non-coding, Nucleic Acids Res. 36 (2008) D170–D172.
[17] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[18] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment, Bioinformatics (Oxford, Engl.) 25 (2009) 1966–1967.
[19] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, D. Marshall, Tablet – next generation sequence assembly visualization, Bioinformatics (Oxford, Engl.) 26 (2010) 401–402.
[20] S. Moxon, F. Schwach, T. Dalmay, D. Maclean, D.J. Studholme, V. Moulton, A toolkit for analysing large-scale plant small RNA datasets, Bioinformatics (Oxford, Engl.) 24 (2008) 2252–2253.
[21] B.H. Zhang, X.P. Pan, S.B. Cox, G.P. Cobb, T.A. Anderson, Evidence that miRNAs are different from other RNAs, CMLS 63 (2006) 246–254.
[22] G. Pertea, X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, J. Quackenbush, TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, Bioinformatics (Oxford, Engl.) 19 (2003) 651–652.
[23] X. Huang, A. Madan, CAP3: A DNA sequence assembly program, Genome Res. 9 (1999) 868–877.
[24] X. Dai, P.X. Zhao, psRNATarget: a plant small RNA target analysis server, Nucleic Acids Res. 39 (2011) W155–W159.
[25] A. Conesa, S. Götz, Blast2GO: a comprehensive suite for functional analysis in plant genomics, Int. J. Plant Genom. 2008 (2008) 619832.
[26] F.R. Kulcheski, F.C. Marcelino-Guimaraes, A.L. Nepomuceno, R.V. Abdelnoor, R. Margis, The use of microRNAs as reference genes for quantitative polymerase chain reaction in soybean, Anal. Biochem. 406 (2010) 185–192.
[27] F.R. Kulcheski, L.F. de Oliveira, L.G. Molina, M.P. Almerao, F.A. Rodrigues, J. Marcolino, J.F. Barbosa, R. Stolf-Moreira, A.L. Nepomuceno, F.C. Marcelino-Guimaraes, R.V. Abdelnoor, L.C. Nascimento, M.F. Carazzolle, G.A. Pereira, R. Margis, Identification of novel soybean microRNAs involved in abiotic and biotic stresses, BMC Genom. 12 (2011) 307.
[28] C. Chen, D.A. Ridzon, A.J. Broomer, Z. Zhou, D.H. Lee, J.T. Nguyen, M. Barbisin, N.L. Xu, V.R. Mahuvakar, M.R. Andersen, K.Q. Lao, K.J. Livak, K.J. Guegler, Real-time quantification of microRNAs by stem-loop RT-PCR, Nucleic Acids Res. 33 (2005) e179.
[29] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Genome Biol. 3 (2002), research0034.
[30] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) Method, Methods 25 (2001) 402–408.
[31] R. Sunkar, G. Jagadeeswaran, In silico identification of conserved microRNAs in large number of diverse plant species, BMC Plant Biol. 8 (2008) 37.
[32] T.P. Frazier, F. Xie, A. Freistaedter, C.E. Burklew, B. Zhang, Identification and characterization of microRNAs and their target genes in tobacco (*Nicotiana tabacum*), Planta 232 (2010) 1289–1308.
[33] T. Unver, I. Parmaksiz, E. Dündar, Identification of conserved micro-RNAs and their target transcripts in opium poppy (*Papaver somniferum* L.), Plant Cell Rep. 29 (2010) 757–769.
[34] F. Guzman, M.P. Almerao, A.P. Korbes, G. Loss-Morais, R. Margis, Identification of microRNAs from Eugenia uniflora by high-throughput sequencing and bioinformatics analysis, PLoS ONE 7 (2012) e49811.
[35] A.P. Korbes, R.D. Machado, F. Guzman, M.P. Almerao, L.F. de Oliveira, G. Loss-Morais, A.C. Turchetto-Zolet, A. Cagliari, F. dos Santos Maraschin, M. Margis-Pinheiro, R. Margis, Identifying conserved and novel microRNAs in developing seeds of Brassica napus using deep sequencing, PLoS ONE 7 (2012) e50663.
[36] Q.-X. Song, Y.-F. Liu, X.-Y. Hu, W.-K. Zhang, B. Ma, S.-Y. Chen, J.-S. Zhang, Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing, BMC Plant Biol. 11 (2011) 5.
[37] R.D. Morin, G. Aksay, E. Dolgosheina, H.A. Ebhardt, V. Magrini, E.R. Mardis, S.C. Sahinalp, P.J. Unrau, Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*, Genome Res. 18 (2008) 571–584.
[38] L. Guo, Z. Lu, Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data, Comput. Biol. Chem. 34 (2010) 165–171.
[39] L. Guo, H. Li, T. Liang, J. Lu, Q. Yang, Q. Ge, Z. Lu, Consistent isomiR expression patterns and 3′ addition events in miRNA gene clusters and families implicate functional and evolutionary relationships, Mol. Biol. Rep. 39 (2012) 6699–6706.
[40] R. Schwab, J.F. Palatnik, M. Riester, C. Schommer, M. Schmid, D. Weigel, Specific effects of microRNAs on the plant transcriptome, Dev. Cell 8 (2005) 517–527.
[41] B. Zhang, X. Pan, E.J. Stellwag, Identification of soybean microRNAs and their targets, Planta 229 (2008) 161–182.
[42] G. Szittya, S. Moxon, D.M. Santos, R. Jing, M.P.S. Fevereiro, V. Moulton, T. Dalmay, High-throughput sequencing of *Medicago truncatula* short RNAs identifies eight new miRNA families, BMC Genom. 9 (2008) 593.
[43] C.-Z. Zhao, H. Xia, T.P. Frazier, Y.-Y. Yao, Y.-P. Bi, A.-Q. Li, M.-J. Li, C.-S. Li, B.-H. Zhang, X.-J. Wang, Deep sequencing identifies novel and conserved microRNAs in peanuts (*Arachis hypogaea* L.), BMC Plant Biol. 10 (2010) 3.
[44] Q.-H. Zhu, A. Spriggs, L. Matthew, L. Fan, G. Kennedy, F. Gubler, C. Helliwell, A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains, Genome Res. 18 (2008) 1456–1465.
[45] H. Vaucheret, Post-transcriptional small RNA pathways in plants: mechanisms and regulations, Genes Develop. 20 (2006) 759–771.
[46] R. Rajagopalan, H. Vaucheret, J. Trejo, D.P. Bartel, A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*, Genes Develop. 20 (2006) 3407–3425.
[47] A.J. Herr, Pathways through the small RNA world of plants, FEBS Lett. 579 (2005) 5879–5888.
[48] F. Vazquez, *Arabidopsis* endogenous small RNAs: highways and byways, Trends Plant Sci. 11 (2006) 460–468.
[49] F.C. Favoreto, C.R. Carvalho, A.B.P. Lima, A. Ferreira, W.R. Clarindo, Genome size and base composition of Bromeliaceae species assessed by flow cytometry, Plant Syst. Evol. 298 (2012) 1185–1193.
[50] D. Gonzalez-Ibeas, J. Blanca, L. Donaire, M. Saladié, A. Mascarell-Creus, A. Cano-Delgado, J. Garcia-Mas, C. Llave, M.A. Aranda, Analysis of the melon (*Cucumis melo*) small RNAome by high-throughput pyrosequencing, BMC Genom. 12 (2011) 393.
[51] J.R. Puzey, A. Karger, M. Axtell, E.M. Kramer, Deep annotation of *Populus trichocarpa* microRNAs from diverse tissue sets, PLoS ONE 7 (2012) e33034.
[52] S. Lv, X. Nie, L. Wang, X. Du, S.S. Biradar, X. Jia, S. Weining, Identification and Characterization of MicroRNAs from Barley (*Hordeum vulgare* L.) by High-Throughput Sequencing, Int. J. Mol. Sci. 13 (2012) 2973–2984.
[53] N. Fahlgren, M.D. Howell, K.D. Kasschau, E.J. Chapman, C.M. Sullivan, J.S. Cumbie, S.A. Givan, T.F. Law, S.R. Grant, J.L. Dangl, J.C. Carrington, High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of MIRNA genes, PLoS ONE 2 (2007) e219.
[54] G. Martínez, J. Forment, C. Llave, V. Pallás, G. Gómez, High-throughput sequencing, characterization and detection of new and conserved cucumber miRNAs, PLoS ONE 6 (2011) e19523.
[55] C. Wang, X. Wang, N.K. Kibet, C. Song, C. Zhang, X. Li, J. Han, J. Fang, Deep sequencing of grapevine flower and berry short RNA library for discovery of novel microRNAs and validation of precise sequences of grapevine microRNAs deposited in miRBase, Physiol. Plant. 143 (2011) 64–81.
[56] V. Pantaleo, G. Szittya, S. Moxon, L. Miozzi, V. Moulton, T. Dalmay, J. Burgyan, Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis, Plant J. Cell. Mol. Biol. 62 (2010) 960–976.
[57] D. De Paola, F. Cattonaro, D. Pignone, G. Sonnante, The miRNAome of globe artichoke: conserved and novel micro RNAs and target analysis, BMC Genom. 13 (2012) 41.
[58] M. Colaiacovo, A. Subacchi, P. Bagnaresi, A. Lamontanara, L. Cattivelli, P. Faccioli, A computational-based update on microRNAs and their targets in barley (*Hordeum vulgare* L.), BMC Genom. 11 (2010) 595.
[59] G. Wu, M.Y. Park, S.R. Conway, J.-W. Wang, D. Weigel, R.S. Poethig, The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*, Cell 138 (2009) 750–759.
[60] M. Shikata, T. Koyama, N. Mitsuda, M. Ohme-Takagi, *Arabidopsis* SBP-box genes SPL10, SPL11 and SPL2 control morphological change in association with shoot maturation in the reproductive phase, Plant Cell Physiol. 50 (2009) 2133–2145.
[61] M.-F. Wu, Q. Tian, J.W. Reed, *Arabidopsis* microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction, Development (Cambr., Engl.) 133 (2006) 4211–4218.
[62] J.H. Yang, S.J. Han, E.K. Yoon, W.S. Lee, Evidence of an auxin signal pathway, microRNA167-ARF8-GH3, and its response to exogenous auxin in cultured rice cells, Nucleic Acids Res. 34 (2006) 1892–1899.
[63] S. Fujii, I. Small, The evolution of RNA editing and pentatricopeptide repeat genes, New Phytol. 191 (2011) 37–47.
[64] T. Eulgem, P.J. Rushton, S. Robatzek, I.E. Somssich, The WRKY superfamily of plant transcription factors, Trends Plant Sci. 5 (2000) 199–206.
[65] H. Banno, Y. Ikeda, Q.W. Niu, N.H. Chua, Overexpression of *Arabidopsis* ESR1 induces initiation of shoot regeneration, Plant Cell 13 (2001) 2609–2619.

[66] T. Kirch, R. Simon, M. Grunewald, W. Werr, The Dornroschen/Enhancer of shoot regeneration1 gene of *Arabidopsis* acts in the control of meristem cell fate and lateral organ development, Plant Cell 15 (2003) 694–705.
[67] C.A. Airoldi, F.D. Rovere, G. Falasca, G. Marino, M. Kooiker, M.M. Altamura, S. Citterio, M.M. Kater, The *Arabidopsis* BET bromodomain factor GTE4 is involved in maintenance of the mitotic cell cycle during plant development, Plant Physiol. 152 (2010) 1320–1334.
[68] S. Zhang, Y. Yue, L. Sheng, Y. Wu, G. Fan, A. Li, X. Hu, M. ShangGuan, C. Wei, PASmiR: a literature-curated database for miRNA molecular regulation in plant response to abiotic stress, BMC Plant Biol. 13 (2013) 1–8.

Anexo 2

Outras produções cientificas relacionadas no período

ARTICLE

doi:10.1038/nature10158

Genome sequence and analysis of the tuber crop potato

The Potato Genome Sequencing Consortium*

Potato (*Solanum tuberosum* L.) is the world's most important non-grain food crop and is central to global food security. It is clonally propagated, highly heterozygous, autotetraploid, and suffers acute inbreeding depression. Here we use a homozygous doubled-monoploid potato clone to sequence and assemble 86% of the 844-megabase genome. We predict 39,031 protein-coding genes and present evidence for at least two genome duplication events indicative of a palaeopolyploid origin. As the first genome sequence of an asterid, the potato genome reveals 2,642 genes specific to this large angiosperm clade. We also sequenced a heterozygous diploid clone and show that gene presence/absence variants and other potentially deleterious mutations occur frequently and are a likely cause of inbreeding depression. Gene family expansion, tissue-specific expression and recruitment of genes to new pathways contributed to the evolution of tuber development. The potato genome sequence provides a platform for genetic improvement of this vital crop.

Potato (*Solanum tuberosum* L.) is a member of the Solanaceae, an economically important family that includes tomato, pepper, aubergine (eggplant), petunia and tobacco. Potato belongs to the asterid clade of eudicot plants that represents ~25% of flowering plant species and from which a complete genome sequence has not yet, to our knowledge, been published. Potato occupies a wide eco-geographical range[1] and is unique among the major world food crops in producing stolons (underground stems) that under suitable environmental conditions swell to form tubers. Its worldwide importance, especially within the developing world, is growing rapidly, with production in 2009 reaching 330 million tons (http://www.fao.org). The tubers are a globally important dietary source of starch, protein, antioxidants and vitamins[2], serving the plant as both a storage organ and a vegetative propagation system. Despite the importance of tubers, the evolutionary and developmental mechanisms of their initiation and growth remain elusive.

Outside of its natural range in South America, the cultivated potato is considered to have a narrow genetic base resulting originally from limited germplasm introductions to Europe. Most potato cultivars are autotetraploid ($2n = 4x = 48$), highly heterozygous, suffer acute inbreeding depression, and are susceptible to many devastating pests and pathogens, as exemplified by the Irish potato famine in the mid-nineteenth century. Together, these attributes present a significant barrier to potato improvement using classical breeding approaches. A challenge to the scientific community is to obtain a genome sequence that will ultimately facilitate advances in breeding.

To overcome the key issue of heterozygosity and allow us to generate a high-quality draft potato genome sequence, we used a unique homozygous form of potato called a doubled monoploid, derived using classical tissue culture techniques[3]. The draft genome sequence from this genotype, *S. tuberosum* group Phureja DM1-3 516 R44 (hereafter referred to as DM), was used to integrate sequence data from a heterozygous diploid breeding line, *S. tuberosum* group Tuberosum RH89-039-16 (hereafter referred to as RH). These two genotypes represent a sample of potato genomic diversity; DM with its fingerling (elongated) tubers was derived from a primitive South American cultivar whereas RH more closely resembles commercially cultivated tetraploid potato. The combined data resources, allied to deep transcriptome sequence from both genotypes, allowed us to explore potato genome structure and organization, as well as key aspects of the biology and evolution of this important crop.

Genome assembly and annotation

We sequenced the nuclear and organellar genomes of DM using a whole-genome shotgun sequencing (WGS) approach. We generated 96.6 Gb of raw sequence from two next-generation sequencing (NGS) platforms, Illumina Genome Analyser and Roche Pyrosequencing, as well as conventional Sanger sequencing technologies. The genome was assembled using SOAPdenovo[4], resulting in a final assembly of 727 Mb, of which 93.9% is non-gapped sequence. Ninety per cent of the assembly falls into 443 superscaffolds larger than 349 kb. The 17-nucleotide depth distribution (Supplementary Fig. 1) suggests a genome size of 844 Mb, consistent with estimates from flow cytometry[5]. Our assembly of 727 Mb is 117 Mb less than the estimated genome size. Analysis of the DM scaffolds indicates 62.2% repetitive content in the assembled section of the DM genome, less than the 74.8% estimated from bacterial artificial chromosome (BAC) and fosmid end sequences (Supplementary Table 1), indicating that much of the unassembled genome is composed of repetitive sequences.

We assessed the quality of the WGS assembly through alignment to Sanger-derived phase 2 BAC sequences. In an alignment length of ~1 Mb (99.4% coverage), no gross assembly errors were detected (Supplementary Table 2 and Supplementary Fig. 2). Alignment of fosmid and BAC paired-end sequences to the WGS scaffolds revealed limited (≤0.12%) potential misassemblies (Supplementary Table 3). Extensive coverage of the potato genome in this assembly was confirmed using available expressed sequence tag (EST) data; 97.1% of 181,558 available Sanger-sequenced *S. tuberosum* ESTs (>200 bp) were detected. Repetitive sequences account for at least 62.2% of the assembled genome (452.5 Mb) (Supplementary Table 1) with long terminal repeat retrotransposons comprising the majority of the transposable element classes, representing 29.4% of the genome. In addition, subtelomeric repeats were identified at or near chromosomal ends (Fig. 1). Using a newly constructed genetic map based on 2,603 polymorphic markers in conjunction with other available

*Lists of authors and their affiliations appear at the end of the paper.

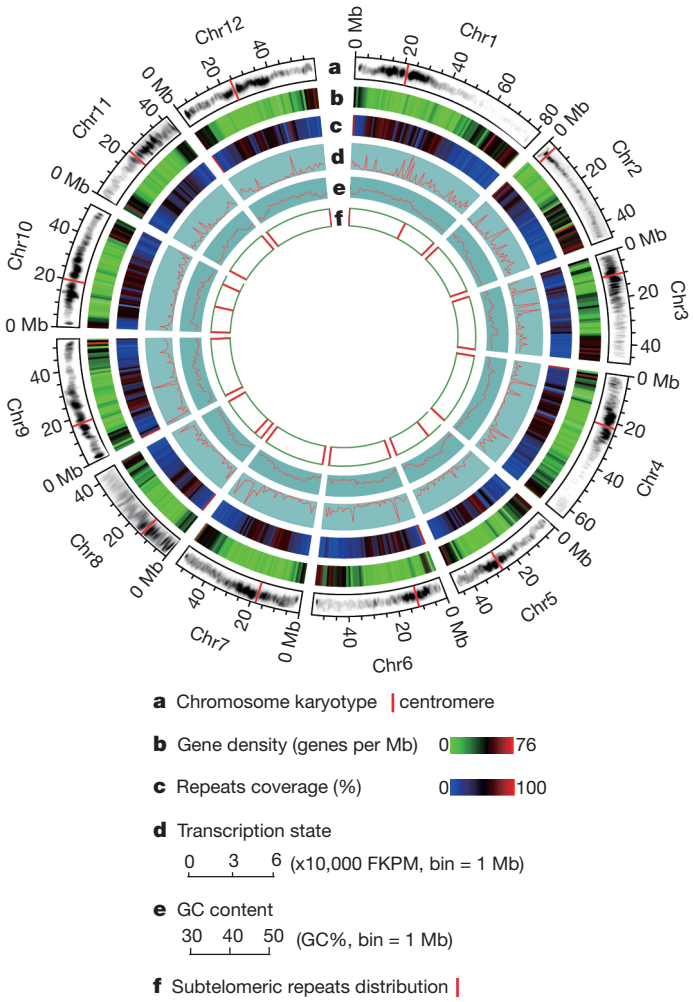

Figure 1 | The potato genome. **a**, Ideograms of the 12 pseudochromosomes of potato (in Mb scales). Each of the 12 pachytene chromosomes from DM was digitally aligned with the ideogram (the amount of DNA in each unit of the pachytene chromosomes is not in proportion to the scales of the pseudochromosomes). **b**, Gene density represented as number of genes per Mb (non-overlapping, window size = 1 Mb). **c**, Percentage of coverage of repetitive sequences (non-overlapping windows, window size = 1 Mb). **d**, Transcription state. The transcription level for each gene was estimated by averaging the fragments per kb exon model per million mapped reads (FPKM) from different tissues in non-overlapping 1-Mb windows. **e**, GC content was estimated by the per cent G+C in 1-Mb non-overlapping windows. **f**, Distribution of the subtelomeric repeat sequence CL14_cons.

genetic and physical maps, we genetically anchored 623 Mb (86%) of the assembled genome (Supplementary Fig. 3), and constructed pseudomolecules for each of the 12 chromosomes (Fig. 1), which harbour 90.3% of the predicted genes.

To aid annotation and address a series of biological questions, we generated 31.5 Gb of RNA-Seq data from 32 DM and 16 RH libraries representing all major tissue types, developmental stages and responses to abiotic and biotic stresses (Supplementary Table 4). For annotation, reads were mapped against the DM genome sequence (90.2% of 824,621,408 DM reads and 88.6% of 140,375,647 RH reads) and in combination with *ab initio* gene prediction, protein and EST alignments, we annotated 39,031 protein-coding genes. RNA-Seq data revealed alternative splicing; 9,875 genes (25.3%) encoded two or more isoforms, indicative of more functional variation than represented by the gene set alone. Overall, 87.9% of the gene models were supported by transcript and/or protein similarity with only 12.1% derived solely from *ab initio* gene predictions (Supplementary Table 5).

Karyotypes of RH and DM suggested similar heterochromatin content[6] (Supplementary Table 6 and Supplementary Fig. 4) with large blocks of heterochromatin located at the pericentromeric regions (Fig. 1). As observed in other plant genomes, there was an inverse relationship between gene density and repetitive sequences (Fig. 1). However, many predicted genes in heterochromatic regions are expressed, consistent with observations in tomato[7] that genic 'islands' are present in the heterochromatic 'ocean'.

Genome evolution

Potato is the first sequenced genome of an asterid, a clade within eudicots that encompasses nearly 70,000 species characterized by unique morphological, developmental and compositional features[8]. Orthologous clustering of the predicted potato proteome with 11 other green plant genomes revealed 4,479 potato genes in 3,181 families in common (Fig. 2a); 24,051 potato genes clustered with at least one of the 11 genomes. Filtering against transposable elements and 153 nonasterid and 57 asterid publicly available transcript-sequence data sets yielded 2,642 high-confidence asterid-specific and 3,372 potato-lineage-specific genes (Supplementary Fig. 5); both sets were enriched for genes of unknown function that had less expression support than the core Viridiplantae genes. Genes encoding transcription factors, self-incompatibility, and defence-related proteins were evident in the asterid-specific gene set (Supplementary Table 7) and presumably contribute to the unique characteristics of asterids.

Structurally, we identified 1,811 syntenic gene blocks involving 10,046 genes in the potato genome (Supplementary Table 8). On the basis of these pairwise paralogous segments, we calculated an age distribution based on the number of transversions at fourfold degenerate sites (4DTv) for all duplicate pairs. In general, two significant groups of blocks are seen in the potato genome (4DTv ~0.36 and ~1.0; Fig. 2b), suggesting two whole-genome duplication (WGD) events. We also identified collinear blocks between potato and three rosid genomes (*Vitis vinifera*, *Arabidopsis thaliana* and *Populus trichocarpa*) that also suggest both events (Fig. 2c and Supplementary Fig. 6). The ancient WGD corresponds to the ancestral hexaploidization (γ) event in grape (Fig. 2b), consistent with a previous report based on EST analysis that the two main branches of eudicots, the asterids and rosids, may share the same palaeo-hexaploid duplication event[9]. The γ event probably occurred after the divergence between dicots and monocots about 185 ± 55 million years ago[10]. The recent duplication can therefore be placed at ~67 million years ago, consistent with the WGD that occurred near the Cretaceous–Tertiary boundary (~65 million years ago)[11]. The divergence of potato and grape occurred at ~89 million years ago (4DTv ~0.48), which is likely to represent the split between the rosids and asterids.

Haplotype diversity

High heterozygosity and inbreeding depression are inherent to potato, a species that predominantly outcrosses and propagates by means of vegetative organs. Indeed, the phenotypes of DM and RH differ, with RH more vigorous than DM (Fig. 3a). To explore the extent of haplotype diversity and possible causes of inbreeding depression, we sequenced and assembled 1,644 RH BAC clones generating 178 Mb of non-redundant sequence from both haplotypes (~10% of the RH genome with uneven coverage) (Supplementary Tables 9–11). After filtering to remove repetitive sequences, we aligned 99 Mb of RH sequence (55%) to the DM genome. These regions were largely collinear with an overall sequence identity of 97.5%, corresponding to one single-nucleotide polymorphism (SNP) every 40 bp and one insertion/deletion (indel) every 394 bp (average length 12.8 bp). Between the two RH haplotypes, 6.6 Mb of sequence could be aligned with 96.5% identity, corresponding to 1 SNP per 29 bp and 1 indel per 253 bp (average length 10.4 bp).

Current algorithms are of limited use in *de novo* whole-genome assembly or haplotype reconstruction of highly heterozygous genomes

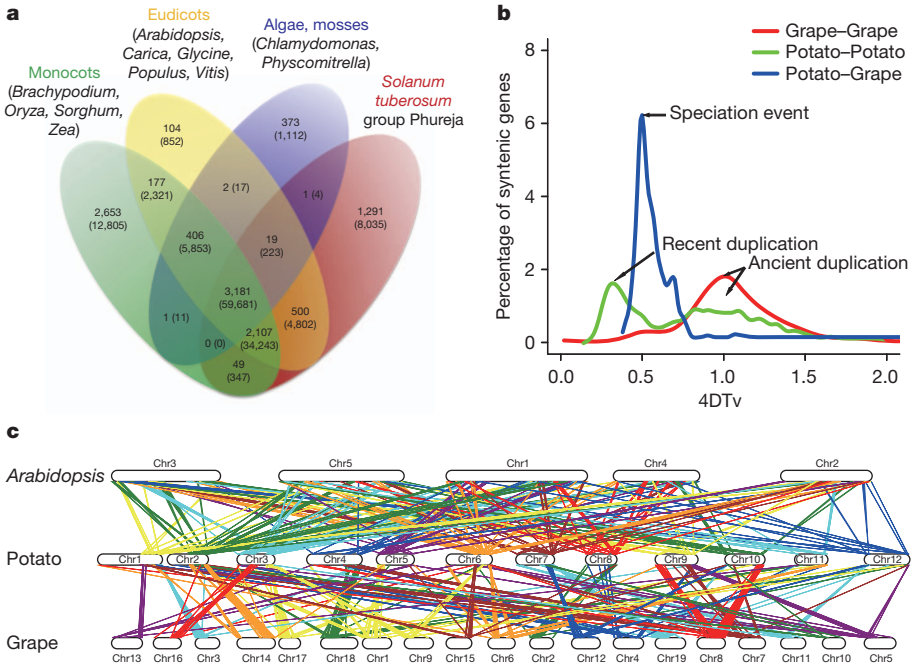

Figure 2 | Comparative analyses and evolution of the potato genome. **a**, Clusters of orthologous and paralogous gene families in 12 plant species as identified by OrthoMCL[33]. Gene family number is listed in each of the components; the number of genes within the families for all of the species within the component is noted within parentheses. **b**, Genome duplication in dicot genomes as revealed through 4DTv analyses. **c**, Syntenic blocks between *A. thaliana*, potato, and *V. vinifera* (grape) demonstrating a high degree of conserved gene order between these taxa.

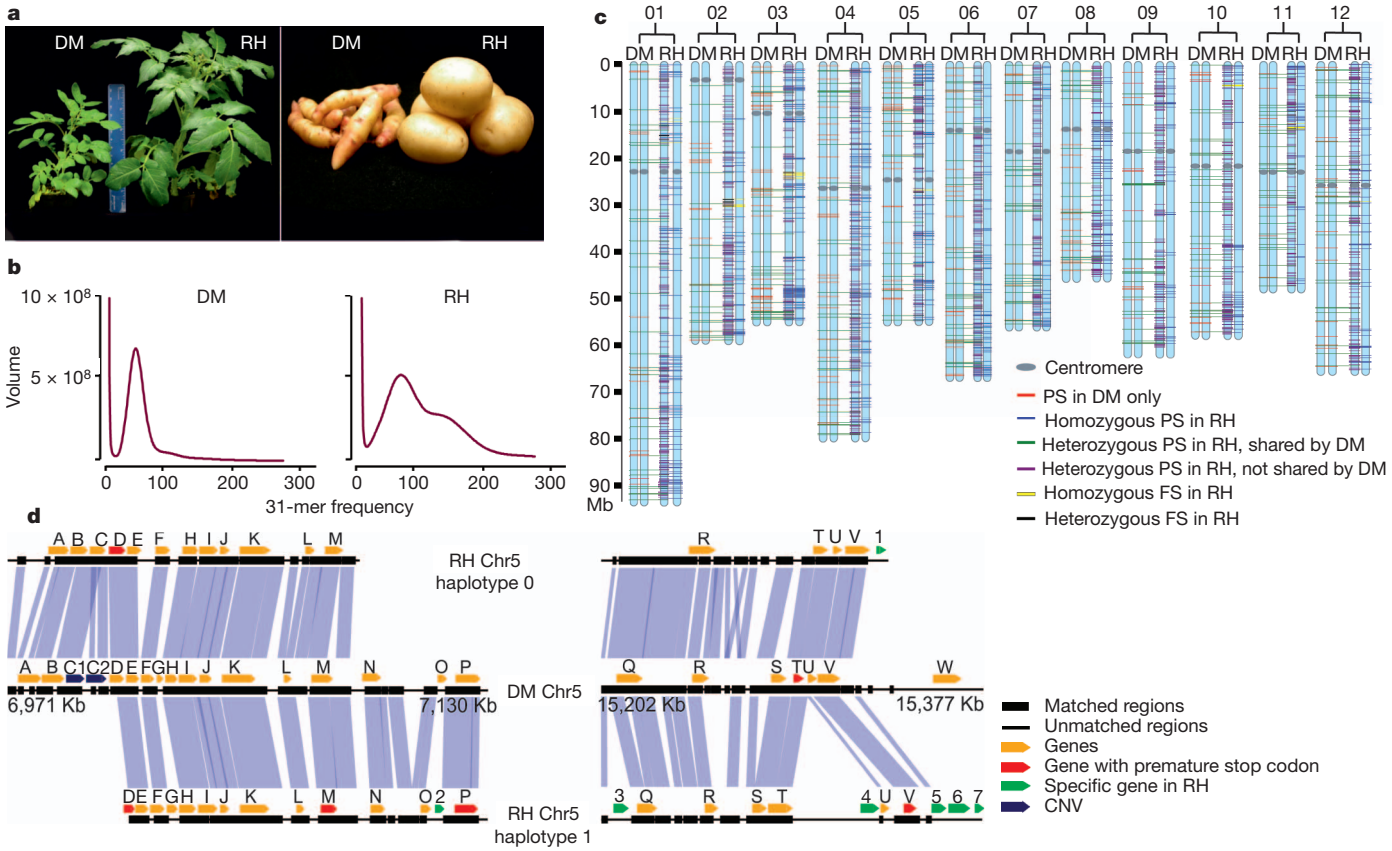

Figure 3 | Haplotype diversity and inbreeding depression. **a**, Plants and tubers of DM and RH showing that RH has greater vigour. **b**, Illumina K-mer volume histograms of DM and RH. The volume of K-mers (*y*-axis) is plotted against the frequency at which they occur (*x*-axis). The leftmost truncated peaks at low frequency and high volume represent K-mers containing essentially random sequencing errors, whereas the distribution to the right represents proper (putatively error-free) data. In contrast to the single modality of DM, RH exhibits clear bi-modality caused by heterozygosity. **c**, Genomic distribution of premature stop, frameshift and presence/absence variation mutations contributing to inbreeding depression. The hypothetical RH pseudomolecules were solely inferred from the corresponding DM ones. Owing to the inability to assign heterozygous PS and FS of RH to a definite haplotype, all heterozygous PS and FS were arbitrarily mapped to the left haplotype of RH. **d**, A zoom-in comparative view of the DM and RH genomes. The left and right alignments are derived from the euchromatic and heterochromatic regions of chromosome 5, respectively. Most of the gene annotations, including PS and RH-specific genes, are supported by transcript data.

such as RH, as shown by K-mer frequency count histograms (Fig. 3b and Supplementary Table 12). To complement the BAC-level comparative analysis and provide a genome-wide perspective of heterozygosity in RH, we mapped 1,118 million whole-genome NGS reads from RH (84× coverage) onto the DM assembly. A total of 457.3 million reads uniquely aligned providing 90.6% (659.1 Mb) coverage. We identified 3.67 million SNPs between DM and one or both haplotypes of RH, with an error rate of 0.91% based on evaluation of RH BAC sequences. We used this data set to explore the possible causes of inbreeding depression by quantifying the occurrence of premature stop, frameshift and presence/absence variants[12], as these disable gene function and contribute to genetic load (Supplementary Tables 13–16). We identified 3,018 SNPs predicted to induce premature stop codons in RH, with 606 homozygous (in both haplotypes) and 2,412 heterozygous. In DM, 940 premature stop codons were identified. In the 2,412 heterozygous RH premature stop codons, 652 were shared with DM and the remaining 1,760 were found in RH only (Fig. 3c and Supplementary Table 13). Frameshift mutations were identified in 80 loci within RH, 49 homozygous and 31 heterozygous, concentrated in seven genomic regions (Fig. 3c and Supplementary Table 14). Finally, we identified presence/absence variations for 275 genes; 246 were RH specific (absent in DM) and 29 were DM specific, with 125 and 9 supported by RNA-Seq and/or Gene Ontology[13] annotation for RH and DM, respectively (Supplementary Tables 15 and 16). Collectively, these data indicate that the complement of homozygous deleterious alleles in DM may be responsible for its reduced level of vigour (Fig. 3a).

The divergence between potato haplotypes is similar to that reported between out-crossing maize accessions[14] and, coupled with our inability to successfully align 45% of the BAC sequences, intra- and inter-genome diversity seem to be a significant feature of the potato genome. A detailed comparison of the three haplotypes (DM and the two haplotypes of RH) at two genomic regions (334 kb in length) using the RH BAC sequence (Fig. 3d and Supplementary Tables 17 and 18) revealed considerable sequence and structural variation. In one region ('euchromatic'; Fig. 3d) we observed one instance of copy number variation, five genes with premature stop codons, and seven RH-specific genes. These observations indicate that the plasticity of the potato genome is greater than revealed from the unassembled RH NGS. Improved assembly algorithms, increased read lengths, and *de novo* sequences of additional haplotypes will reveal the full catalogue of genes critical to inbreeding depression.

Tuber biology

In developing DM and RH tubers, 15,235 genes were expressed in the transition from stolons to tubers, with 1,217 transcripts exhibiting >5-fold expression in stolons versus five RH tuber tissues (young tuber, mature tuber, tuber peel, cortex and pith; Supplementary Table 19). Of these, 333 transcripts were upregulated during the transition from stolon to tuber, with the most highly upregulated transcripts encoding storage proteins. Foremost among these were the genes encoding proteinase inhibitors and patatin (15 genes), in which the phospholipase A function has been largely replaced by a protein storage function in the tuber[15]. In particular, a large family of 28 Kunitz protease inhibitor genes (KTIs) was identified with twice the number of genes in potato compared to tomato. The KTI genes are distributed across the genome with individual members exhibiting specific expression patterns (Fig. 4a, b). KTIs are frequently induced after pest and pathogen attack and act primarily as inhibitors of exogenous proteinases[16]; therefore the expansion of the KTI family may provide resistance to biotic stress for the newly evolved vulnerable underground organ.

The stolon to tuber transition also coincides with strong upregulation of genes associated with starch biosynthesis (Fig. 4c). We observed several starch biosynthetic genes that were 3–8-fold more highly expressed in tuber tissues of RH compared to DM (Fig. 4c). Together this suggests a stronger shift from the relatively low sink strength of the ATP-generating general carbon metabolism reactions towards the plastidic starch synthesis pathway in tubers of RH, thereby causing a flux of carbon into the amyloplast. This contrasts with the cereal endosperm where carbon is transported into the amyloplast in the form of ADP-glucose via a specific transporter (brittle 1 protein[17]). Carbon transport into the amyloplasts of potato tubers is primarily in the form of glucose-6-phosphate[18], although recent evidence indicates that glucose-1-phosphate is quantitatively important under certain conditions[19]. The transport mechanism for glucose-1-phosphate is unknown and the genome sequence contains six genes for hexose-phosphate transporters with two highly and specifically expressed in stolons and tubers. Furthermore, an additional 23 genes encode proteins homologous to other carbohydrate derivative transporters, such as triose phosphate, phosphoenolpyruvate, or UDP-glucuronic acid transporters and two loci with homologues for the brittle 1 protein. By contrast, in leaves, carbon-fixation-specific genes such as plastidic aldolase, fructose-1,6-biphosphatase and distinct leaf isoforms of starch synthase, starch branching enzyme, starch phosphorylase and ADP-glucose pyrophosphorylase were upregulated. Of particular interest is the difference in tuber expression of enzymes involved in the hydrolytic and phosphorolytic starch degradation pathways. Considerably greater levels of α-amylase (10–25-fold) and β-amylase (5–10-fold) mRNAs were found in DM tubers compared to RH, whereas α-1,4 glucan phosphorylase mRNA was equivalent in DM and RH tubers. These gene expression differences between the breeding line RH and the more primitive DM are consistent with the concept that increasing tuber yield may be partially attained by selection for decreased activity of the hydrolytic starch degradation pathway.

Recent studies using a potato genotype strictly dependent on short days for tuber induction (*S. tuberosum* group Andigena) identified a potato homologue (*SP6A*) of *A. thaliana FLOWERING LOCUS T* (*FT*) as the long-distance tuberization inductive signal. SP6A is produced in the leaves, consistent with its role as the mobile signal (S. Prat, personal communication). *SP/FT* is a multi-gene family (Supplementary Text and Supplementary Fig. 7) and expression of a second *FT* homologue, *SP5G*, in mature tubers suggests a possible function in the control of tuber sprouting, a photoperiod-dependent phenomenon[20]. Likewise, expression of a homologue of the *A. thaliana* flowering time MADS box gene *SOC1*, acting downstream of *FT*[21], is restricted to tuber sprouts (Supplementary Fig. 8). Expression of a third *FT* homologue, *SP3D*, does not correlate with tuberization induction but instead with transition to flowering, which is regulated independently of day length (S. Prat, personal communication). These data indicate that neofunctionalization of the day-length-dependent flowering control pathway has occurred in potato to control formation and possibly sprouting of a novel storage organ, the tuber (Supplementary Fig. 9).

Disease resistance

Potato is susceptible to a wide range of pests and pathogens and the identification of genes conferring disease resistance has been a major focus of the research community. Most cloned disease resistance genes in the Solanaceae encode nucleotide-binding site (NBS) and leucine-rich-repeat (LRR) domains. The DM assembly contains 408 NBS-LRR-encoding genes, 57 Toll/interleukin-1 receptor/plant R gene homology (TIR) domains and 351 non-TIR types (Supplementary Table 20), similar to the 402 resistance (*R*) gene candidates in *Populus*[22]. Highly related homologues of the cloned potato late blight resistance genes *R1*, *RB*, *R2*, *R3a*, *Rpi-blb2* and *Rpi-vnt1.1* were present in the assembly. In RH, the chromosome 5 R1 cluster contains two distinct haplotypes; one is collinear with the R1 region in DM (Supplementary Fig. 10), yet neither the DM nor the RH R1 regions are collinear with other potato R1 regions[23,24]. Comparison of the DM potato *R* gene sequences with well-established gene models (functional *R* genes) indicates that many NBS-LRR genes (39.4%) are pseudogenes owing to indels, frameshift mutations, or premature stop

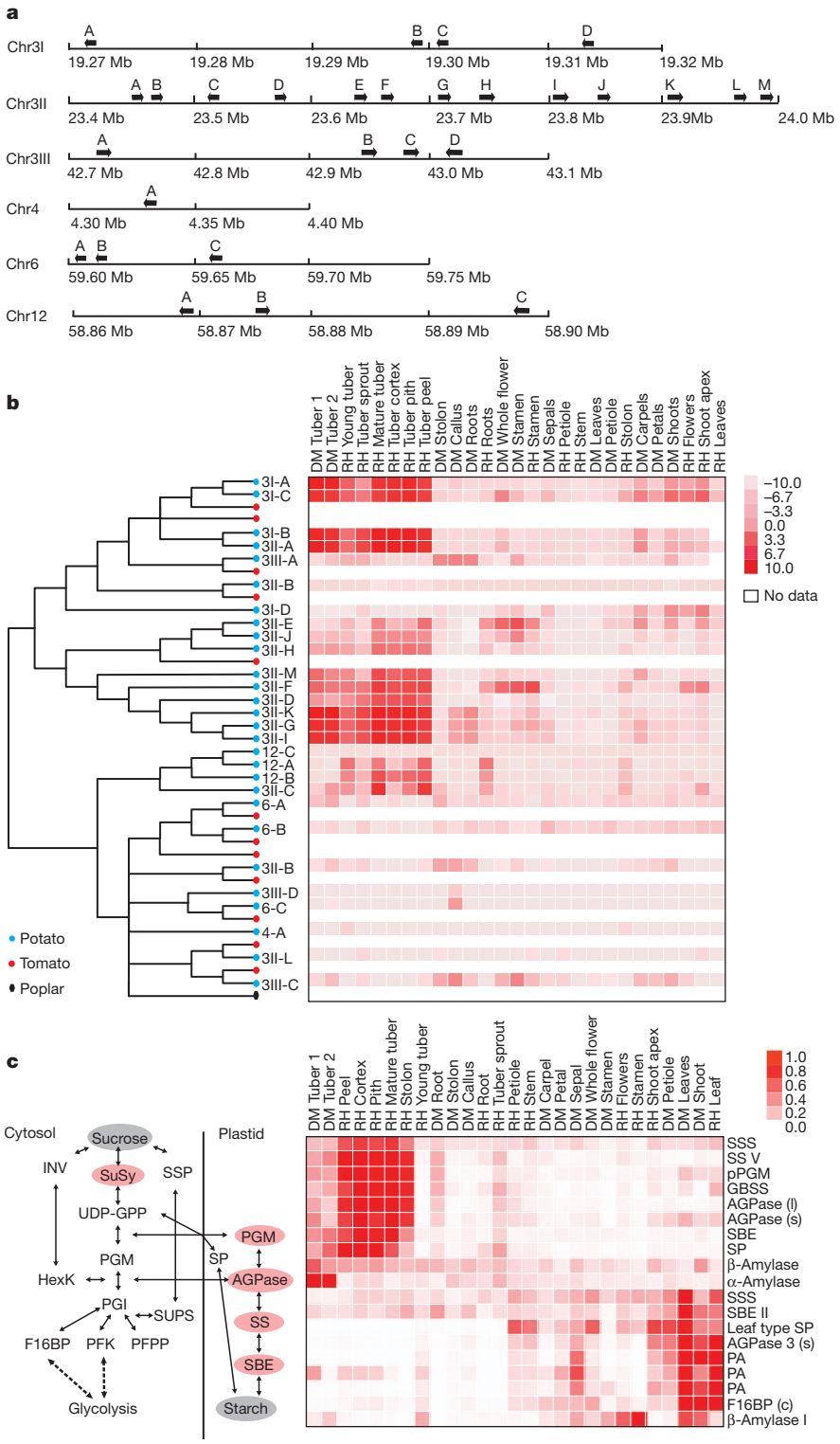

Figure 4 | Gene expression of selected tissues and genes. **a**, KTI gene organization across the potato genome. Black arrows indicate the location of individual genes on six scaffolds located on four chromosomes. **b**, Phylogenetic tree and KTI gene expression heat map. The KTI genes were clustered using all potato and tomato genes available with the *Populus* KTI gene as an out-group. The tissue specificity of individual members of the highly expanded potato gene family is shown in the heat map. Expression levels are indicated by shades of red, where white indicates no expression or lack of data for tomato and poplar. **c**, A model of starch synthesis showing enzyme activities is shown on the left. AGPase, ADP-glucose pyrophosphorylase; F16BP, fructose-1,6-biphosphatase; HexK, hexokinase; INV, invertase; PFK, phosphofructokinase; PFPP, pyrophosphate-fructose-6-phosphate-1-phosphotransferase; PGI, phosphoglucose isomerase; PGM, phosphoglucomutase; SBE, starch branching enzyme; SP, starch phosphorylase; SPP, sucrose phosphate phosphatase; SS, starch synthase; SuSy, sucrose synthase; SUPS, sucrose phosphate synthase; UDP-GPP, UDP-glucose pyrophosphorylase. The grey background denotes substrate (sucrose) and product (starch) and the red background indicates genes that are specifically upregulated in RH versus DM. On the right, a heat map of the genes involved in carbohydrate metabolism is shown. ADP-glucose pyrophosphorylase large subunit, AGPase (l); ADP-glucose pyrophosphorylase small subunit, AGPase (s); ADP-glucose pyrophosphorylase small subunit 3, AGPase 3 (s); cytosolic fructose-1,6-biphosphatase, F16BP (c); granule bound starch synthase, GBSS; leaf type L starch phosphorylase, Leaf type SP; plastidic phosphoglucomutase, pPGM; starch branching enzyme II, SBE II; soluble starch synthase, SSS; starch synthase V, SSV; three variants of plastidic aldolase, PA.

codons including the *R1*, *R3a* and *Rpi-vnt1.1* clusters that contain extensive chimaeras and exhibit evolutionary patterns of type I *R* genes[25]. This high rate of pseudogenization parallels the rapid evolution of effector genes observed in the potato late blight pathogen, *Phytophthora infestans*[26]. Coupled with abundant haplotype diversity, tetraploid potato may therefore contain thousands of *R*-gene analogues.

Conclusions and future directions

We sequenced a unique doubled-monoploid potato clone to overcome the problems associated with genome assembly due to high levels of heterozygosity and were able to generate a high-quality draft potato genome sequence that provides new insights into eudicot genome evolution. Using a combination of data from the vigorous, heterozygous diploid RH and relatively weak, doubled-monoploid DM, we could directly address the form and extent of heterozygosity in potato and provide the first view into the complexities that underlie inbreeding depression. Combined with other recent studies, the potato genome sequence may elucidate the evolution of tuberization. This evolutionary innovation evolved exclusively in the *Solanum* section *Petota* that encompasses ~200 species distributed from the southwestern United States to central Argentina and Chile. Neighbouring *Solanum* species,

including the *Lycopersicon* section, which comprises wild and cultivated tomatoes, did not acquire this trait. Both gene family expansion and recruitment of existing genes for new pathways contributed to the evolution of tuber development in potato.

Given the pivotal role of potato in world food production and security, the potato genome provides a new resource for use in breeding. Many traits of interest to plant breeders are quantitative in nature and the genome sequence will simplify both their characterization and deployment in cultivars. Whereas much genetic research is conducted at the diploid level in potato, almost all potato cultivars are tetraploid and most breeding is conducted in tetraploid material. Hence, the development of experimental and computational methods for routine and informative high-resolution genetic characterization of polyploids remains an important goal for the realization of many of the potential benefits of the potato genome sequence.

METHODS SUMMARY

DM1-3 516 R44 (DM) resulted from chromosome doubling of a monoploid ($1n = 1x = 12$) derived by anther culture of a heterozygous diploid ($2n = 2x = 24$) *S. tuberosum* group Phureja clone (PI 225669)[27]. RH89-039-16 (RH) is a diploid clone derived from a cross between a *S. tuberosum* 'dihaploid' (SUH2293) and a diploid clone (BC1034) generated from a cross between two *S. tuberosum* × *S. tuberosum* group Phureja hybrids[28] (Supplementary Fig. 11). Sequence data from three platforms, Sanger, Roche 454 Pyrosequencing, and Illumina Sequencing-by-Synthesis, were used to assemble the DM genome using the SOAPdenovo assembly algorithm[4]. The RH genotype was sequenced using shotgun sequencing of BACs and WGS in which reads were mapped to the DM reference assembly. Superscaffolds were anchored to the 12 linkage groups using a combination of *in silico* and genetic mapping data. Repeat sequences were identified through sequence similarity at the nucleotide and protein level[29]. Genes were annotated using a combined approach[30] on the repeat masked genome with *ab initio* gene predictions, protein similarity and transcripts to build optimal gene models. Illumina RNA-Seq reads were mapped to the DM draft sequence using Tophat[31] and expression levels from the representative transcript were determined using Cufflinks[32].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 11 January; accepted 3 May 2011.
Published online 10 July 2011.

1. Hijmans, R. J. Global distribution of the potato crop. *Am. J. Potato Res.* **78,** 403–412 (2001).
2. Burlingame, B., Mouillé, B. & Charrondiére, R. Nutrients, bioactive non-nutrients and anti-nutrients in potatoes. *J. Food Compost. Anal.* **22,** 494–502 (2009).
3. Paz, M. M. & Veilleux, R. E. Influence of culture medium and *in vitro* conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids. *Plant Breed.* **118,** 53–57 (1999).
4. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–272 (2010).
5. Arumuganathan, K. & Earle, E. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9,** 208–218 (1991).
6. Tang, X. *et al.* Assignment of genetic linkage maps to diploid *Solanum tuberosum* pachytene chromosomes by BAC-FISH technology. *Chromosome Res.* **17,** 899–915 (2009).
7. Peters, S. A. *et al. Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *Plant J.* **58,** 857–869 (2009).
8. Albach, D. C., Soltis, P. S. & Soltis, D. E. Patterns of embryological and biochemical evolution in the Asterids. *Syst. Bot.* **26,** 242–262 (2001).
9. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18,** 1944–1954 (2008).
10. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
11. Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl Acad. Sci. USA* **106,** 5737–5742 (2009).
12. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genet.* **42,** 1027–1030 (2010).
13. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
14. Gore, M. A. *et al.* A first-generation haplotype map of maize. *Science* **326,** 1115–1117 (2009).
15. Prat, S. *et al.* Gene expression during tuber development in potato plants. *FEBS Lett.* **268,** 334–338 (1990).
16. Glaczinski, H., Heibges, A., Salamini, R. & Gebhardt, C. Members of the Kunitz-type protease inhibitor gene family of potato inhibit soluble tuber invertase *in vitro*. *Potato Res.* **45,** 163–176 (2002).
17. Shannon, J. C., Pien, F. M. & Liu, K. C. Nucleotides and nucleotide sugars in developing maize endosperms: synthesis of ADP-glucose in *brittle-1*. *Plant Physiol.* **110,** 835–843 (1996).
18. Tauberger, E. *et al.* Antisense inhibition of plastidial phosphoglucomutase provides compelling evidence that potato tuber amyloplasts import carbon from the cytosol in the form of glucose-6-phosphate. *Plant J.* **23,** 43–53 (2000).
19. Fettke, J. *et al.* Glucose 1-phosphate is efficiently taken up by potato (*Solanum tuberosum*) tuber parenchyma cells and converted to reserve starch granules. *New Phytol.* **185,** 663–675 (2010).
20. Sonnewald, U. Control of potato tuber sprouting. *Trends Plant Sci.* **6,** 333–335 (2001).
21. Yoo, S. K. *et al. CONSTANS* activates *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* through *FLOWERING LOCUS T* to promote flowering in *Arabidopsis*. *Plant Physiol.* **139,** 770–778 (2005).
22. Kohler, A. *et al.* Genome-wide identi?cation of *NBS* resistance genes in *Populus trichocarpa*. *Plant Mol. Biol.* **66,** 619–636 (2008).
23. Ballvora, A. *et al.* Comparative sequence analysis of *Solanum* and *Arabidopsis* in a hot spot for pathogen resistance on potato chromosome V reveals a patchwork of conserved and rapidly evolving genome segments. *BMC Genomics* **8,** 112 (2007).
24. Kuang, H. *et al.* The *R1* resistance gene cluster contains three groups of independently evolving, type I *R1* homologues and shows substantial structural variation among haplotypes of *Solanum demissum*. *Plant J.* **44,** 37–51 (2005).
25. Kuang, H., Woo, S. S., Meyers, B. C., Nevo, E. & Michelmore, R. W. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell* **16,** 2870–2894 (2004).
26. Haas, B. J. *et al.* Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461,** 393–398 (2009).
27. Haynes, F. L. in *Prospects for the Potato in the Developing World: an International Symposium on Key Problems and Potentials for Greater Use of the Potato in the Developing World* (ed. French, E. R.) 100–110 (International Potato Center (CIP), 1972).
28. van Os, H. *et al.* Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* **173,** 1075–1087 (2006).
29. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25,** 4.10.1–4.10.14 (2004).
30. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8,** R13 (2007).
31. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
32. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
33. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the assistance of W. Amoros, B. Babinska, R. V. Baslerov, B. K. Bumazhkin, M. F. Carboni, T. Conner, J. Coombs, L. Daddiego, J. M. D'Ambrosio, G. Diretto, S. B. Divito, D. Douches, M. Filipiak, G. Gianese, R. Hutten, E. Jacobsen, E. Kalinska, S. Kamoun, D. Kells, H. Kossowska, L. Lopez, M. Magallanes-Lundback, T. Miranda, P. S. Naik, A. N. Panteleeva, D. Pattanayak, E. O. Patutina, M. Portantier, S. Rawat, R. Simon, B. P. Singh, B. Singh, W. Stiekema, M. V. Sukhacheva and C. Town in providing plant material, generating data, annotation, analyses, and discussions. We are indebted to additional faculty and staff of the BGI-Shenzhen, J Craig Venter Institute, and MSU Research Technology Support Facility who contributed to this project. Background and preliminary data were provided by the Centre for BioSystems Genomics (CBSG), EU-project (APOPHYS EU-QLRT-2001-01849) and US Department of Agriculture National Institute of Food and Agriculture SolCAP project (2008-55300-04757 and 2009-85606-05673). We acknowledge the funding made available by the "863" National High Tech Research Development Program in China (2006AA100107), "973" National Key Basic Research Program in China (2006CB101904, 2007CB815703, 2007CB815705, 2009CB119000), Board of Wageningen University and Research Centre, CAPES - Brazilian Ministry of Education, Chinese Academy of Agricultural Sciences (seed grant to S.H.), Chinese Ministry of Agriculture (The "948" Program), Chinese Ministry of Finance (1251610601001), Chinese Ministry of Science and Technology (2007DFB30080), China Postdoctoral Science Foundation (20070420446 to Z.Z.), CONICET (Argentina), DAFF Research Stimulus Fund (07-567), CONICYT-Chile (PBCT-PSD-03), Danish Council for Strategic Research Programme Commission on Health, Food and Welfare (2101-07-0116), Danish Council for Strategic Research Programme Commission on Strategic Growth Technologies (Grant 2106-07-0021), FINCyT ((099-FINCyT-EQUIP-2009)/(076-FINCyT-PIN-2008), Préstamo BID no. 1663/OC-PE, FONDAP and BASAL-CMM), Fund for Economic Structural Support (FES), HarvestPlus Challenge Program, Indian Council of Agricultural Research, INIA-Ministry of Agriculture of Chile, Instituto Nacional de Innovacion Agraria-Ministry of Agriculture of Peru, Instituto Nacional de Tecnología Agropecuaria (INTA), Italian Ministry of Research (Special Fund for Basic Research), International Potato Center (CIP-CGIAR core funds), LBMG of Center for Genome Regulation and Center for Mathematical Modeling, Universidad de Chile (UMI 2807 CNRS), Ministry of Education and Science of Russia (contract 02.552.11.7073), National Nature Science Foundation of China (30671319, 30725008, 30890032, 30971995), Natural Science Foundation of Shandong Province in China (Y2006D21), Netherlands Technology Foundation (STW), Netherlands Genomics Initiative (NGI), Netherlands Ministries of Economic Affairs (EZ) and Agriculture (LNV), New Zealand Institute for Crop & Food Research Ltd

Strategic Science Initiative, Perez Guerrero Fund, Peruvian Ministry of Agriculture-Technical Secretariat of coordination with the CGIAR, Peruvian National Council of Science and Technology (CONCYTEC), Polish Ministry of Science and Higher Education (47/PGS/2006/01), Programa Cooperativo para el Desarrollo Tecnológico Agroalimentario y Agroindustrial del Cono Sur (PROCISUR), Project Programa Bicentenario de Ciencia y Tecnología - Conicyt, PBCT - Conicyt PSD-03, Russian Foundation for Basic Research (09-04-12275), Secretaría de Ciencia y Tecnología (SECyT) actual Ministerio de Ciencia y Tecnología (MINCyT), Argentina, Shenzhen Municipal Government of China (CXB200903110066A, ZYC200903240077A, ZYC200903240076A), Solexa project (272-07-0196), Special Multilateral Fund of the Inter-American Council for Integral Development (FEMCIDI), Teagasc, Teagasc Walsh Fellowship Scheme, The New Zealand Institute for Plant & Food Research Ltd Capability Fund, UK Potato Genome Sequencing grant (Scottish Government Rural and Environment Research and Analysis Directorate (RERAD), Department for Environment, Food and Rural Affairs (DEFRA), Agriculture and Horticulture Development Board - Potato Council), UK Biotechnology and Biological Sciences Research Council (Grant BB/F012640), US National Science Foundation Plant Genome Research Program (DBI-0604907 /DBI-0834044), Virginia Agricultural Experiment Station USDA Hatch Funds (135853), and Wellcome Trust Strategic award (WT 083481).

Author Contributions A.D.G., A.G., A.N.M., A.V.B., A.V.M., B.B.K., B.K., B.R.W., B.S., B.t.L.H., B.V., B.X., B.Z., C.L., C.R.B., C.W.B.B., D.F.M., D. Martinez, D. Milbourne, D.M.A.M., D.M.B., D.D., D.M., E.D., F.G., G.A.M., G.A.T., G.D.I.C., G.G., G.J. Bishop, G.J. Bryan, G.L., G.O., G.P., G.Z., H.K., H.L., H.v.E., I.N., J.d.B., J.G., J.H., J.J., J.M.E.J., J.W., J.X., K.L.N., K.O'B., L.D., L.E.B., M.B., M.D., M.d.R.H., M.F., M. Geoffroy, M.Ghislain, M.I., M.P., M.S., M.T., N.M., N.V.R., O.P., P.F., P.N., P.S., Q.H., R.C.H.J.v.H., R.E.V., R.G., R.G.F.V., R. Lozano, R. Li, S.C., S.E.F., S.H., S.J.T., S.K.C., S.K.S., S.L., S.P., S.Y., T.B., T.V.K., V.U.P., X. Xiong, X. Xu, Y.D., Y.H., Y.L., Y.Y., Y.Z. and Z.Z. were involved in experimental design, data generation and/or data analysis. A.N.M., B.K., C.R.B., C.W.B.B., D.D., D. Milbourne, D.M.A.M., D.M.B., E.D., G.G., G.J. Bishop, G.J. Bryan, G.O., H.L., I.N., J.d.B., J.J., J.M.E.J., K.L.N., M.B., M.F., M.D., M.S., O.P., R.C.H.J.v.H., R.E.V., R.G.F.V., R. Lozano, R.W., S.E.F., S.H., S.J.T., S.K.S., T.B. and X. Xu wrote the manuscript. B.S., C.R.B., C.W.B.B., D.F.M., D. Milbourne, D.M.A.M., D.Q., G.G., G.J. Bishop, G.J. Bryan, G.O., G.P., J.M.E.J., J.W., K.G.S., R.G.F.V., R. Li, R.W., S.E.F., S.H., S.K.C., S.Y., W.Z. and Y.D. supervised data generation/analysis and managed the project. C.R.B., C.W.B.B., G.J. Bryan, G.O., J.M.E.J. and S.H. are members of The Potato Genome Sequencing Consortium Steering Committee.

Author Information BAC and fosmid end sequences have been deposited in the GSS division of GenBank (BAC: GS025503–GS026177, GS262924–GS365942, GS504213–GS557003; fosmid: FI900795–FI901529, FI907952–FI927051, GS557234–GS594339, GS635316–GS765761). DM Illumina GA2 WGS and Roche 454 sequences have been deposited in the NCBI Sequence Read Archive (SRA029323) and EBI Short Read Archive (ERP000411) respectively. RH NGS sequences have been deposited in the EBI Short Read Archive (ERP000627). DM and RH RNA-Seq reads have been deposited in the NCBI Sequence Read Archive (SRA030516; study SRP005965) and the European Nucleotide Database ArrayExpress Database (E-MTAB-552; study ERP000527), respectively. The DM Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AEWC00000000. The version described in this paper is the First Version, AEWC01000000. Genome sequence and annotation can be obtained and viewed at http://potatogenome.net. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.H. (huangsanwen@caas.net.cn), C.R.B. (buell@msu.edu) or R.G.F.V. (Richard.Visser@wur.nl).

The Potato Genome Consortium (Participants are listed alphabetically by institution.)

BGI-Shenzhen Xun Xu[1], Shengkai Pan[1], Shifeng Cheng[1], Bo Zhang[1], Desheng Mu[1], Peixiang Ni[1], Gengyun Zhang[1], Shuang Yang (Principal Investigator)[1], Ruiqiang Li (Principal Investigator)[1], Jun Wang (Principal Investigator)[1]; **Cayetano Heredia University** Gisella Orjeda (Principal Investigator)[2], Frank Guzman[2], Michael Torres[2], Roberto Lozano[2], Olga Ponce[2], Diana Martinez[2], Germán De la Cruz[2]; **Central Potato Research Institute** S. K. Chakrabarti (Principal Investigator)[3], Virupaksh U. Patil[3]; **Centre Bioengineering RAS** Konstantin G. Skryabin (Principal Investigator)[4], Boris B. Kuznetsov[4], Nikolai V. Ravin[4], Tatjana V. Kolganova[4], Alexey V. Beletsky[4], Andrei V. Mardanov[4]; **CGR-CMM, Universidad de Chile** Alex Di Genova[5]; **College of Life Sciences, University of Dundee** Daniel M. Bolser[6], David M. A. Martin (Principal Investigator)[6]; **High Technology Research Center, Shandong Academy of Agricultural Sciences** Guangcun Li[7], Yu Yang[7]; **Huazhong Agriculture University** Hanhui Kuang[8], Qun Hu[8]; **Hunan Agricultural University** Xingyao Xiong[9]; **Imperial College London** Gerard J. Bishop[10]; **Instituto de Investigaciones Agropecuarias** Boris Sagredo (Principal Investigator)[11], Nilo Mejía[11]; **Institute of Biochemistry & Biophysics** Wlodzimierz Zagorski (Principal Investigator)[12], Robert Gromadka[12], Jan Gawor[12], Pawel Szczesny[12]; **Institute of Vegetables & Flowers, Chinese Academy of Agricultural Sciences** Sanwen Huang (Principal Investigator)[13], Zhonghua Zhang[13], Chunbo Liang[13], Jun He[13], Ying Li[13], Ying He[13], Jianfei Xu[13], Youjun Zhang[13], Binyan Xie[13], Yongchen Du[13], Dongyu Qu (Principal Investigator)[13]; **International Potato Center** Merideth Bonierbale[14], Marc Ghislain[14], Maria del Rosario Herrera[14]; **Italian National Agency for New Technologies, Energy & Sustainable Development** Giovanni Giuliano (Principal Investigator)[15], Marco Pietrella[15], Gaetano Perrotta[15], Paolo Facella[15]; **J Craig Venter Institute** Kimberly O'Brien[16]; **Laboratorio de Agrobiotecnología, Instituto Nacional de Tecnología Agropecuaria** Sergio E. Feingold (Principal Investigator)[17], Leandro E. Barreiro[17], Gabriela A. Massa[17]; **Laboratorio de Biología de Sistemas, Universidad Nacional de LaPlata** Luis Diambra[18]; **Michigan State University** Brett R. Whitty[19], Brieanne Vaillancourt[19], Haining Lin[19], Alicia N. Massa[19], Michael Geoffroy[19], Steven Lundback[19], Dean DellaPenna[19], C. Robin Buell (Principal Investigator)[19]; **Scottish Crop Research Institute** Sanjeev Kumar Sharma[20]†, David F. Marshall[20]†, Robbie Waugh[20]†, Glenn J. Bryan (Principal Investigator)[20]†; **Teagasc Crops Research Centre** Marialaura Destefanis[21], Istvan Nagy[21], Dan Milbourne (Principal Investigator)[21]; **The New Zealand Institute for Plant & Food Research Ltd** Susan J. Thomson[22], Mark Fiers[22], Jeanne M. E. Jacobs (Principal Investigator)[22]; **University of Aalborg** Kåre L. Nielsen (Principal Investigator)[23], Mads Sønderkær[23]; **University of Wisconsin** Marina Iovene[24], Giovana A. Torres[24], Jiming Jiang (Principal Investigator)[24]; **Virginia Polytechnic Institute & State University** Richard E. Veilleux[25]; **Wageningen University & Research Centre** Christian W. B. Bachem (Principal Investigator)[26], Jan de Boer[26], Theo Borm[26], Bjorn Kloosterman[26], Herman van Eck[26], Erwin Datema[27], Bas te Lintel Hekkert[27], Aska Goverse[28,29], Roeland C. H. J. van Ham[27,28] & Richard G. F. Visser[26,28]

[1]BGI-Shenzhen, Chinese Ministry of Agricultural, Key Lab of Genomics, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China. [2]Cayetano Heredia University, Genomics Research Unit, Av Honorio Delgado 430, Lima 31, Peru and San Cristobal of Huamanga University, Biotechnology and Plant Genetics Laboratory, Ayacucho, Peru. [3]Central Potato Research Institute, Shimla 171001, Himachal Pradesh, India. [4]Centre Bioengineering RAS, Prospekt 60-letya Oktyabrya, 7-1, Moscow 117312, Russia. [5]Center for Genome Regulation and Center for Mathematical Modeling, Universidad de Chile (UMI 2807 CNRS), Chile. [6]College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK. [7]High Technology Research Center, Shandong Academy of Agricultural Sciences, 11 Sangyuan Road, Jinan 250100, P. R. China. [8]Huazhong Agriculture University, Ministry of Education, College of Horticulture and Forestry, Department of Vegetable Crops, Key Laboratory of Horticulture Biology, Wuhan 430070, P. R. China. [9]Hunan Agricultural University, College of Horticulture and Landscape, Changsha, Hunan 410128, China. [10]Imperial College London, Division of Biology, South Kensington Campus, London SW7 1AZ, UK. [11]Instituto de Investigaciones Agropecuarias, Avda. Salamanca s/n, Km 105 ruta 5 sur, sector Los Choapinos. Rengo, Región del Libertador Bernardo O'Higgins, Código Postal 2940000, Chile. [12]Institute of Biochemistry and Biophysics, DNA Sequencing and Oligonucleotides Synthesis Laboratory, PAS ul. Pawinskiego 5a, 02-106 Warsaw, Poland. [13]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Key Laboratory of Horticultural Crops Genetic Improvement of Ministry of Agriculture, Sino-Dutch Joint Lab of Horticultural Genomics Technology, Beijing 100081, China. [14]International Potato Center, P.O. Box 1558, Lima 12, Peru. [15]Italian National Agency for New Technologies, Energy and Sustainable Development (ENEA), Casaccia Research Center, Via Anguillarese 301, 00123 Roma, Italy and Trisaia Research Center, S.S. 106 Ionica - Km 419,50 75026 Rotondella (Matera), Italy. [16]J Craig Venter Institute, 9712 Medical Center Dr, Rockville, Maryland 20850, USA. [17]Laboratorio de Agrobiotecnología, Estación Experimental Agropecuaria Balcarce, Instituto Nacional de Tecnología Agropecuaria (INTA) cc276 (7620) Balcarce, Argentina. [18]Laboratorio de Biología de Sistemas, CREG, Universidad Nacional de LaPlata, 1888, Argentina. [19]Michigan State University, East Lansing, Michigan 48824, USA. [20]Scottish Crop Research Institute, Genetics Programme, Invergowrie, Dundee DD2 5DA, UK. [21]Teagasc Crops Research Centre, Oak Park, Carlow, Ireland. [22]The New Zealand Institute for Plant & Food Research Ltd., Private Bag 4704, Christchurch 8140, New Zealand. [23]University of Aalborg (AAU), Department of Biotechnology, Chemistry and Environmental Engineering, Sohngaardsholmsvej 49, 9000 Aalborg, Denmark. [24]University of Wisconsin-Madison, Department of Horticulture, 1575 Linden Drive, Madison, Wisconsin 53706, USA. [25]Virginia Polytechnic Institute and State University, Department of Horticulture, 544 Latham Hall, Blacksburg, Virginia 24061, USA. [26]Wageningen University and Research Centre, Dept. of Plant Sciences, Laboratory of Plant Breeding, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. [27]Wageningen University and Research Centre, Applied Bioinformatics, Plant Research International, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. [28]Centre for BioSystems Genomics, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. [29]Wageningen University and Research Centre, Dept. of Plant Sciences, Laboratory of Nematology, Droevendaalsesteeg 1, 6708PB Wageningen, Netherlands. †Present address: The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK (S.K.S., D.F.M., R.W., G.J. Bryan).

METHODS

DM whole-genome shotgun sequencing and assembly. Libraries were constructed from DM genomic DNA and sequenced on the Sanger, Illumina Genome Analyser 2 (GA2) and Roche 454 platforms using standard protocols (see Supplementary Text). A BAC library and three fosmid libraries were end sequenced using the Sanger platform. For the Illumina GA2 platform, we generated 70.6 Gb of 37–73 bp paired-end reads from 16 libraries with insert lengths of 200–811 bp (Supplementary Tables 21 and 22). We also generated 18.7 Gb of Illumina mate-pair libraries (2, 5 and 10 kb insert size). In total, 7.2 Gb of 454 single-end data were generated and applied for gap filling to improve the assembly, of which 4.7 Gb (12,594,513 reads) were incorporated into the final assembly. For the 8 and 20 kb 454 paired-end reads, representing 0.7 and 1.0 Gb of raw data respectively, 90.7 Mb (511,254 reads) and 211 Mb (1,525,992 reads), respectively, were incorporated into the final assembly.

We generated a high-quality potato genome using the short read assembly software SOAPdenovo[4] (Version 1014). We first assembled 69.4 Gb of GA2 paired-end short reads into contigs, which are sequence assemblies without gaps composed of overlapping reads. To increase the assembly accuracy, only 78.3% of the reads with high quality were considered. Then contigs were further linked into scaffolds by paired-end relationships (~300 to ~550 bp insert size), mate-pair reads (2 to approximately 10 kb), fosmid ends (~40 kb, 90,407 pairs of end sequences) and BAC ends (~100 kb, 71,375 pairs of end sequences). We then filled gaps with the entire short-read data generated using Illumina GA2 reads. The primary contig N_{50} size (the contig length such that using equal or longer contigs produces half of the bases of the assembled genome) was 697 bp and increased to 1,318 kb after gap-filling (Supplementary Tables 23 and 24). When only the paired-end relationships were used in the assembly process, the N_{50} scaffold size was 22.4 kb. Adding mate-pair reads with 2, 5 and 10 kb insert sizes, the N_{50} scaffold size increased to 67, 173 and 389 kb, respectively. When integrated with additional libraries of larger insert size, such as fosmid and BAC end sequences, the N_{50} reached 1,318 kb. The final assembly size was 727 Mb, 93.87% of which is non-gapped sequence. We further filled the gaps with 6.74 fold coverage of 454 data, which increased the N_{50} contig size to 31,429 bp with 15.4% of the gaps filled.

The single-base accuracy of the assembly was estimated by the depth and proportion of disconcordant reads. For the DM v3.0 assembly, 95.45% of 880 million usable reads could be mapped back to the assembled genome by SOAP 2.20 (ref. 34) using optimal parameters. The read depth was calculated for each genomic location and peak depth for whole genome and the CDS regions are 100 and 105, respectively. Approximately 96% of the assembled sequences had more than 20-fold coverage (Supplementary Fig. 1). The overall GC content of the potato genome is about 34.8% with a positive correlation between GC content and sequencing depth (data not shown). The DM potato should have few heterozygous sites and 93.04% of the sites can be supported by at least 90% reads, suggesting high base quality and accuracy.

RH genome sequencing. Whole-genome sequencing of genotype RH was performed on the Illumina GA2 platform using a variety of fragment sizes and reads lengths resulting in a total of 144 Gb of raw data (Supplementary Table 25). These data were filtered using a custom C program and assembled using SOAPdenovo 1.03 (ref. 4). Additionally, four 20-kb mate-pair libraries were sequenced on a Roche 454 Titanium sequencer, amounting to 581 Mb of raw data (Supplementary Table 26). The resulting sequences were filtered for duplicates using custom Python scripts.

The RH BACs were sequenced using a combination of Sanger and 454 sequencing at various levels of coverage (Supplementary Tables 9–11). Consensus base calling errors in the BAC sequences were corrected using custom Python and C scripts using a similar approach to that described previously[35] (Supplementary Text). Sequence overlaps between BACs within the same physical tiling path were identified using megablast from BLAST 2.2.21 (ref. 36) and merged with megamerger from the EMBOSS 6.1.0 package[37]. Using the same pipeline, several kilobase-sized gaps were closed through alignment of a preliminary RH whole-genome assembly. The resulting non-redundant contigs were scaffolded by mapping the RH whole-genome Illumina and 454 mated sequences against these contigs using SOAPalign 2.20 (ref. 34) and subsequently processing these mapping results with a custom Python script. The scaffolds were then ordered into superscaffolds based on the BAC order in the tiling paths of the FPC map. This procedure removed 25 Mb of redundant sequence, reduced the number of sequence fragments from 17,228 to 3,768, and increased the N50 sequence length from 24 to 144 kb (Supplementary Tables 9 and 10).

Construction of the DM genetic map and anchoring of the genome. To anchor and fully orientate physical contigs along the chromosome, a genetic map was developed *de novo* using sequence-tagged-site (STS) markers comprising simple sequence repeats (SSR), SNPs, and diversity array technology (DArT). SSR and SNP markers were designed directly from assembled sequence scaffolds, whereas polymorphic DArT marker sequences were searched against the scaffolds for high-quality unique matches. A total of 4,836 STS markers including 2,174 DArTs, 2,304 SNPs and 358 SSRs were analysed on 180 progeny clones from a backcross population ((DM × DI) × DI) developed at CIP between DM and DI (CIP no. 703825), a heterozygous diploid *S. tuberosum* group Stenotomum (formerly *S. stenotomum* ssp. *goniocalyx*) landrace clone. The data from 2,603 polymorphic STS markers comprising 1,881 DArTs, 393 SNPs and 329 SSR alleles were analysed using JoinMap 4 (ref. 38) and yielded the expected 12 potato linkage groups. Supplementary Fig. 3 represents the mapping and anchoring of the potato genome, using chromosome 7 as an example.

Anchoring the DM genome was accomplished using direct and indirect approaches. The direct approach employed the ((DM × DI) × DI) linkage map whereby 2,037 of the 2,603 STS markers comprised of 1,402 DArTs, 376 SNPs and 259 SSRs could be uniquely anchored on the DM superscaffolds. This approach anchored ~52% (394 Mb) of the assembly arranged into 334 superscaffolds (Supplementary Table 27 and Supplementary Fig. 3).

RH is the male parent of the mapping population of the ultra-high-density (UHD) linkage map[28] used for construction and genetic anchoring of the physical map using the RHPOTKEY BAC library[39]. The indirect mapping approach exploited *in silico* anchoring using the RH genetic and physical map[28,40], as well as tomato genetic map data from SGN (http://solgenomics.net/). Amplified fragment length polymorphism markers from the RH genetic map were linked to DM sequence scaffolds via BLAST alignment[36] of whole-genome-profiling sequence tags[41] obtained from anchored seed BACs in the RH physical map, or by direct alignment of fully sequenced RH seed BACs to the DM sequence. The combined marker alignments were processed into robust anchor points. The tomato sequence markers from the genetic maps were aligned to the DM assembly using SSAHA2 (ref. 42). Positions of ambiguously anchored superscaffolds were manually checked and corrected. This approach anchored an additional ~32% of the assembly (229 Mb). In 294 cases, the two independent approaches provided direct support for each other, anchoring the same scaffold to the same position on the two maps.

Overall, the two strategies anchored 649 superscaffolds to approximate positions on the genetic map of potato covering a length of 623 Mb. The 623 Mb (~86%) anchored genome includes ~90% of the 39,031 predicted genes. Of the unanchored superscaffolds, 84 were found in the N90 (622 scaffolds greater than 0.25 Mb), constituting 17 Mb of the overall assembly or 2% of the assembled genome. The longest anchored superscaffold is 7 Mb (from chromosome 1) and the longest unanchored superscaffold is 2.5 Mb.

Identification of repetitive sequences. Transposable elements (TEs) in the potato genome assembly were identified at the DNA and protein level. RepeatMasker[29] was applied using Repbase[43] for TE identification at the DNA level. At the protein level, RepeatProteinMask[29,44] was used in a WuBlastX[36] search against the TE protein database to further identify TEs. Overlapping TEs belonging to the same repeat class were collated, and sequences were removed if they overlapped >80% and belonged to different repeat classes.

Gene prediction. To predict genes, we performed *ab initio* predictions on the repeat-masked genome and then integrated the results with spliced alignments of proteins and transcripts to genome sequences using GLEAN[30]. The potato genome was masked by identified repeat sequences longer than 500 bp, except for miniature inverted repeat transposable elements which are usually found near genes or inside introns[45]. The software Augustus[46] and Genscan[47] was used for *ab initio* predictions with parameters trained for *A. thaliana*. For similarity-based gene prediction, we aligned the protein sequences of four sequenced plants (*A. thaliana*, *Carica papaya*, *V. vinifera* and *Oryza sativa*) onto the potato genome using TBLASTN with an *E*-value cut-off of 1×10^{-5}, and then similar genome sequences were aligned against the matching proteins using Genewise[48] for accurately spliced alignments. In EST-based predictions, EST sequences of 11 *Solanum* species were aligned against the potato genome using BLAT (identity ≥0.95, coverage ≥0.90) to generate spliced alignments. All these resources and prediction approaches were combined by GLEAN[30] to build the consensus gene set. To finalize the gene set, we aligned the RNA-Seq from 32 libraries, of which eight were sequenced with both single- and paired-end reads, to the genome using Tophat[31] and the alignments were then used as input for Cufflinks[32] using the default parameters. Gene, transcript and peptide sets were filtered to remove small genes, genes modelled across sequencing gaps, TE-encoding genes, and other incorrect annotations. The final gene set contains 39,031 genes with 56,218 protein-coding transcripts, of which 52,925 nonidentical proteins were retained for analysis.

Transcriptome sequencing. RNA was isolated from many tissues of DM and RH that represent developmental, abiotic stress and biotic stress conditions (Supplementary Table 4 and Supplementary Text). cDNA libraries were constructed (Illumina) and sequenced on an Illumina GA2 in the single- and/or paired-end

mode. To represent the expression of each gene, we selected a representative transcript from each gene model by selecting the longest CDS from each gene. The aligned read data were generated by Tophat[31] and the selected transcripts used as input into Cufflinks[32], a short-read transcript assembler that calculates the fragments per kb per million mapped reads (FPKM) as expression values for each transcript. Cufflinks was run with default settings, with a maximum intron length of 15,000. FPKM values were reported and tabulated for each transcript (Supplementary Table 19).

Comparative genome analyses. Paralogous and orthologous clusters were identified using OrthoMCL[49] using the predicted proteomes of 11 plant species (Supplementary Table 28). After removing 1,602 TE-related genes that were not filtered in earlier annotation steps, asterid-specific and potato-lineage-specific genes were identified using the initial OrthoMCL clustering followed by BLAST searches (*E*-value cut-off of 1×10^{-5}) against assemblies of ESTs available from the PlantGDB project (http://plantgdb.org; 153 nonasterid species and 57 asterid species; Supplementary Fig. 5 and Supplementary Table 29). Analysis of protein domains was performed using the Pfam hmm models identified by InterProScan searches against InterPro (http://www.ebi.ac.uk/interpro). We compared the Pfam domains of the asterid-specific and potato-lineage-specific sets with those that are shared with at least one other nonasterid genome or transcriptome. A Fisher's exact test was used to detect significant differences in Pfam representation between protein sets.

After removing the self and multiple matches, the syntenic blocks ($\geq$5 genes per block) were identified using MCscan[9] and i-adhore 3.0 (ref. 50) based on the aligned protein gene pairs (Supplementary Table 8). For the self-aligned results, each aligned block represents the paralogous segments pair that arose from the genome duplication whereas, for the inter-species alignment results, each aligned block represents the orthologous pair derived from the shared ancestor. We calculated the 4DTv (fourfold degenerate synonymous sites of the third codons) for each gene pair from the aligned block and give a distribution for the 4DTv value to estimate the speciation or WGD event that occurred in evolutionary history.

Identification of disease resistance genes. Predicted open reading frames (ORFs) from the annotation of *S. tuberosum* group Phureja assembly V3 were screened using HMMER V.3 (http://hmmer.janelia.org/software) against the raw hidden Markov model (HMM) corresponding to the Pfam NBS (NB-ARC) family (PF00931). The HMM was downloaded from the Pfam home page (http://pfam.sanger.ac.uk/). The analysis using the raw HMM of the NBS domain resulted in 351 candidates. From these, a high quality protein set ($<1 \times 10^{-60}$) was aligned and used to construct a potato-specific NBS HMM using the module 'hmmbuild'. Using this new potato-specific model, we identified 500 NBS-candidate proteins that were individually analysed. To detect TIR and LRR domains, Pfam HMM searches were used. The raw TIR HMM (PF01582) and LRR 1 HMM (PF00560) were downloaded and compared against the two sets of NBS-encoding amino acid sequences using HMMER V3. Both TIR and LRR domains were validated using NCBI conserved domains and multiple expectation maximization for motif elicitation (MEME)[51]. In the case of LRRs, MEME was also useful to detect the number of repeats of this particular domain in the protein. As previously reported[52], Pfam analysis could not identify the CC motif in the N-terminal region. CC domains were thus analysed using the MARCOIL[53] program with a threshold probability of 90 (ref. 52) and double-checked using paircoil2 (ref. 54) with a *P*-score cut-off of 0.025 (ref. 55). Selected genes ($\pm$1.5 kb) were searched using BLASTX against a reference *R*-gene set[56] to find a well-characterized homologue. The reference set was used to select and annotate as pseudogenes those peptides that had large deletions, insertions, frameshift mutations, or premature stop codons. DNA and protein comparisons were used.

Haplotype diversity analysis. RH reads generated by the Illumina GA2 were mapped onto the DM genome assembly using SOAP2.20 (ref. 34) allowing at most four mismatches and SNPs were called using SOAPsnp. Q20 was used to filter the SNPs owing to sequencing errors. To exclude SNP calling errors caused by incorrect alignments, we excluded adjacent SNPs separated by $<$5 bp. SOAPindel was used to detect the indels between DM and RH. Only indels supported by more than three uniquely mapped reads were retained. Owing to the heterozygosity of RH, the SNPs and indels were classified into heterozygous and homozygous SNPs or indels.

On the basis of the annotated genes in the DM genome assembly, we extracted the SNPs located at coding regions and stop codons. If a homozygous SNP in RH within a coding region induced a premature stop codon, we defined the gene harbouring this SNP as a homozygous premature stop gene in RH. If the SNP inducing a premature stop codon was heterozygous, the gene harbouring this SNP was considered a heterozygous premature stop codon gene in RH. In addition, both categories can be further divided into premature stop codons shared with DM or not shared with DM. As a result, the numbers of premature stop codons are 606 homozygous PS genes in RH, 1,760 heterozygous PS genes in RH but not shared with DM, 288 PS in DM only, and 652 heterozygous premature stop codons in RH and shared by DM.

To identify genes with frameshift mutations in RH, we identified all the genes containing indels of which the length could not be divided by 3. We found 80 genes with frameshift mutations, of which 31 were heterozygous and 49 were homozygous.

To identify DM-specific genes, we mapped all the RH Illumina GA2 reads to the DM genome assembly. If the gene was not mapped to any RH read, it was considered a DM-specific gene. We identified 35 DM-specific genes, 11 of which are supported by similarity to entries in the KEGG database[57]. To identify RH-specific genes, we assembled the RH Illumina GA2 reads that did not map to the DM genome into RH-specific scaffolds. Then, these scaffolds were annotated using the same strategy as for DM. To exclude contamination, we aligned the CDS sequences against the protein set of bacteria with the *E*-value cut-off of 1×10^{-5} using Blastx. CDS sequences with $>$90% identity and $>$90% coverage were considered contaminants and were excluded. In addition, all DM RNA-seq reads were mapped onto the CDS sequences, and CDS sequences with homologous reads were excluded because these genes may be due to incorrect assembly. In total, we predicted 246 RH specific genes, 34 of which are supported by Gene Ontology annotation[17].

34. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
35. Chaisson, M., Pevzner, P. & Tang, H. Fragment assembly with short reads. *Bioinformatics* **20,** 2067–2074 (2004).
36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
37. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).
38. Van Ooijen, J. W. in *JoinMap 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations* (ed. Kyazma, B. V.) (Wageningen, 2006).
39. Borm, T. J. *Construction and Use of a Physical Map of Potato.* PhD thesis, Wageningen Univ. (2008).
40. Visser, R. G. F. *et al.* Sequencing the potato genome: outline and first results to come from the elucidation of the sequence of the world's third most important crop. *Am. J. Potato Res.* **86,** 417–429 (2009).
41. Van der Vossen, E. *et al.* in *Whole Genome Profiling of the Diploid Potato Clone RH89-039-16* (Plant & Animal Genomes XVIII Conference, 2010).
42. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
43. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).
44. Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. *DupMasker*: a tool for annotating primate segmental duplications. *Genome Res.* **18,** 1362–1368 (2008).
45. Kuang, H. *et al.* Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res.* **19,** 42–56 (2009).
46. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32,** W309–W312 (2004).
47. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
48. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
49. Chen, F., Mackey, A. J., Vermunt, J. K. & Roos, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* **2,** e383 (2007).
50. Simillion, C., Janssens, K., Sterck, L. & Van de Peer, Y. i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* **24,** 127–128 (2008).
51. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3,** 21–29 (1995).
52. Mun, J. H., Yu, H. J., Park, S. & Park, B. S. Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa. Mol. Genet. Genomics* **282,** 617–631 (2009).
53. Delorenzi, M. & Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **18,** 617–625 (2002).
54. McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved predictions of coiled coils from sequence. *Bioinformatics* **22,** 356–358 (2006).
55. Porter, B. W. *et al.* Genome-wide analysis of *Carica papaya* reveals a small *NBS* resistance gene family. *Mol. Genet. Genomics* **281,** 609–626 (2009).
56. Sanseverino, W. *et al.* PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res.* **38,** D814–D821 (2010).
57. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277–D280 (2004).

OPEN ACCESS Freely available online

PLOS ONE

Identifying Conserved and Novel MicroRNAs in Developing Seeds of *Brassica napus* Using Deep Sequencing

Ana Paula Körbes[1,2], Ronei Dorneles Machado[2], Frank Guzman[1], Mauricio Pereira Almerão[2], Luiz Felipe Valter de Oliveira[1], Guilherme Loss-Morais[2], Andreia Carina Turchetto-Zolet[1,2], Alexandro Cagliari[1], Felipe dos Santos Maraschin[1,3], Marcia Margis-Pinheiro[1,2], Rogerio Margis[1,2,4]*

1 PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **2** PPGBCM, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **3** Departamento de Botânica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **4** Departamento de Biofísica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil

Abstract

MicroRNAs (miRNAs) are important post-transcriptional regulators of plant development and seed formation. In *Brassica napus*, an important edible oil crop, valuable lipids are synthesized and stored in specific seed tissues during embryogenesis. The miRNA transcriptome of *B. napus* is currently poorly characterized, especially at different seed developmental stages. This work aims to describe the miRNAome of developing seeds of *B. napus* by identifying plant-conserved and novel miRNAs and comparing miRNA abundance in mature versus developing seeds. Members of 59 miRNA families were detected through a computational analysis of a large number of reads obtained from deep sequencing two small RNA and two RNA-seq libraries of (i) pooled immature developing stages and (ii) mature *B. napus* seeds. Among these miRNA families, 17 families are currently known to exist in *B. napus*; additionally 29 families not reported in *B. napus* but conserved in other plant species were identified by alignment with known plant mature miRNAs. Assembled mRNA-seq contigs allowed for a search of putative new precursors and led to the identification of 13 novel miRNA families. Analysis of miRNA population between libraries reveals that several miRNAs and isomiRNAs have different abundance in developing stages compared to mature seeds. The predicted miRNA target genes encode a broad range of proteins related to seed development and energy storage. This work presents a comparative study of the miRNA transcriptome of mature and developing *B. napus* seeds and provides a basis for future research on individual miRNAs and their functions in embryogenesis, seed maturation and lipid accumulation in *B. napus*.

Citation: Körbes AP, Machado RD, Guzman F, Almerão MP, de Oliveira LFV, et al. (2012) Identifying Conserved and Novel MicroRNAs in Developing Seeds of *Brassica napus* Using Deep Sequencing. PLoS ONE 7(11): e50663. doi:10.1371/journal.pone.0050663

Editor: Michael Schubert, Ecole Normale Supérieure de Lyon, France

Received May 24, 2012; **Accepted** October 24, 2012; **Published** November 30, 2012

Funding: This work was supported by CAPES, CNPq, CNPq-Universal 472575/2011-2, Genoprot-CNPq-MCT 559636/2009-1, Agroestruturante-FAPERGS-FINEP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rogerio.margis@ufrgs.br

Introduction

Eukaryotic gene expression is regulated at the transcriptional and post-transcriptional levels. An important post-transcriptional mechanism that has recently been discovered is controlled by endogenous, noncoding small RNAs (sRNAs), primarily small interfering RNAs (siRNAs) and microRNAs (miRNAs) [1–4]. In plants, miRNA genes, called primary miRNAs (pri-miRNAs), are typically encoded in intergenic regions and are transcribed by RNA Polymerase II as long polyadenylated transcripts, similar to protein-coding genes [5]. These primary sequences contain an imperfect stem-loop structure that is recognized by DICER-Like1 (DCL1) for sequential cleavage, which converts the pri-miRNAs into the precursor sequences (pre-miRNAs) that are further processed to generate 18–24 nucleotide (nt)-long sequences called mature miRNAs [6]. The imperfect complementary strand to the most abundant miRNA is often called miRNA*, and both strands are originated from the 5p and 3p arms of the pre-miRNA hairpin structure. These sRNAs play critical roles during plant development, regulating a variety of processes, such as embryogenesis, seed germination, organ formation, and developmental timing and patterning [7–13]. The binding of the miRNA to mRNA targets leads to gene silencing by endonucleolytic cleavage or translational inhibition, depending on the degree of complementarity between the miRNA and its target transcript [14–18].

Brassica napus, known as Oilseed Rape, is the third most important edible oil crop worldwide (www.faostat.fao.org). During embryogenesis, *B. napus* seeds build up storage reserves in specific tissues. The vast majority of these reserves are made up of lipids (40–45%) and proteins (17–26%) that are almost exclusively stored in the cotyledons of the maturing embryo [19]. Biogenesis of oil bodies (lipid-containing structures) begins as early as the heart stage in embryogenesis and lipid accumulation rapidly increases during weeks 4–8 after anthesis [20,21]. These developmental stages are correlated with high synthetic lipid activity and a decline in the expression of genes coding for oil biosynthetic and glycolytic enzymes but not of the genes involved in the later steps of oil accumulation [22].

The number of miRNAs in the miRNA registry database (miRBase release 18) [23] that are known in *B. napus* (48 miRNAs) is considerably small when compared to the model plant *Arabidopsis* (328 miRNAs) or to the crop species soybean (395 miRNAs) and rice (661 miRNAs). *B. napus* (Bna) miRNAs were identified in a few previous studies, primarily through either a computational analysis of known plant miRNAs against the Bna Expressed Sequence Tag (EST) and Genome Survey (GSS) sequences [24] or cloning strategies from whole seedlings or vascular exudates of nutrient-stressed plants [25–29]. These strategies allowed for the identification of Bna-miRNAs that are conserved and highly abundant in many plant species [30,31]. The application of high-throughput sequencing technology in functional studies of small RNAs has been useful in accelerating the discovery of low abundance and species-specific miRNAs in plants under several different growing conditions [32–40]. Recently, [41] nine new Bna-miRNAs were reported using deep sequencing to investigate the Bna-miRNA profiles of seeds from high and low oil-content cultivars in very early embryonic development stages. However, the expression patterns and functions of *B. napus* miRNAs from seed development stages to maturation remain largely unknown.

In the present work, we identified miRNAs that may be involved in stages of the *B. napus* seed development process and in the accumulation of storage reserves. Illumina sequencing of two small RNA libraries of immature and mature stages of *B. napus* seeds were used to characterize the miRNAs. In addition, polyadenylated transcript sequencing (mRNA-seq) libraries were used to identify the pre-miRNAs expressed in the seeds that were unknown to science. A total of 251 mature miRNAs from 59 distinct miRNA families were identified in the computational analysis, from which, 29 families were previously unidentified in *B. napus* but conserved in other plants and 13 families were reported for the first time in plants. Several miRNAs were more abundant in seed development stages than in mature seeds, and putative targets were predicted to encode a broad range of proteins related to seed development and energy storage.

Materials and Methods

Plant Material and Growth Conditions

B. napus (cultivar PFB-2, Embrapa) plants were grown in an open environment (30°S 51°W) from May to October 2010. Flowers were tagged upon opening and developing siliques from different plants were collected in the middle of a light cycle at 7, 14, 21, 28, 42 and 50 days after flower opening (DAF). The seeds were dissected from the siliques, immediately frozen in liquid nitrogen and stored at −80°C.

RNA Isolation and Deep Sequencing

Total RNA was isolated from the seed material using Trizol (Invitrogen, CA, USA) according to the manufacturer's protocol, and the RNA quality was evaluated by electrophoresis on a 1% agarose gel. Total RNA (>10 μg) was sent to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for processing. Two sRNA libraries were constructed; one from mature seeds (50 DAF) and one from an equivalent mixture of RNA from immature seeds at DAF stages 7–42. Briefly, the construction of the small RNA libraries consisted of the following successive steps: acrylamide gel purification of the RNA bands corresponding to the size range 20–30 nt, ligation of the 3p and 5p adapters to the RNA in two separate subsequent steps, each followed by acrylamide gel purification, cDNA synthesis followed by acrylamide gel purification, and a final step of PCR amplification to generate a cDNA colony template library for Illumina sequencing. The polyadenylated transcript sequencing (mRNA-seq) was performed using the following successive steps: Poly(A) purification, cDNA synthesis using Poly(T) primer, shotgun to generate inserts of 500 nt, 3p and 5p adapter ligations, pre-amplification, colony generation and Illumina single-end 100 bases sequencing. The libraries were sequenced by Illumina HiSeq2000.

Data Analysis

The overall procedure for analyzing small libraries is shown in Figure S1. All low quality reads (FASTq value <13) were removed, and 5p and 3p adapter sequences were trimmed using Genome Analyzer Pipeline (Fasteris). The remaining low quality reads with 'n' were removed using PrinSeq script [42]. Sequences shorter than 18 nt and longer than 25 nt were excluded from further analysis. Small RNAs derived from Viridiplantae rRNAs, tRNAs, snRNAs and snoRNAs (from the tRNAdb [43], SILVA rRNA [44], and NONCODE v3.0 [45] databases] and small RNAs derived from Rosales mtRNA and cpRNA [from the NCBI GenBank database (http://ftp.ncbi.nlm.nih.gov)] were identified by mapping with Bowtie v 0.12.7 [46] and excluded from further miRNA predictions and analyses.

After cleaning the data (low quality reads, adapter sequences), the mRNA-seq data from the two libraries were pooled and assembled in contigs using the CLC Genome Workbench version 4.0.2 (CLC bio, Aarhus, Denmark) algorithm for *de novo* sequence assembly using the default parameters (similarity = 0.8, length fraction = 0.5, insertion/deletion cost = 3, mismatch cost = 3). In total, 237,993 contigs were assembled and used as reference for the discovery of pre-miRNA and miRNA sequences.

Identification and Analysis of Conserved and Novel miRNAs

To identify plant-conserved miRNAs, small RNA sequences were aligned with known non-redundant plant mature miRNAs (Viridiplantae) and Brassicaceae precursors that were deposited in the miRBase database (Release 18, November 2011) using Bowtie v 0.12.7. Complete alignment of the sequences was required and zero mismatches were allowed. To search for novel miRNAs, small RNA sequences were matched against assembled mRNA-seq contigs using SOAP2 [47]. The SOAP2 output was filtered with an in-house filter tool to separate candidate sequences as miRNA precursors using an anchoring pattern of one or two blocks of aligned small RNAs with a perfect match. As miRNA precursors have a characteristic hairpin structure, the next step to select candidate sequences was secondary structure analysis by RNAfold using an annotation algorithm from the UEA sRNA toolkit [48]. In addition, perfect stem-loop structures should have the miRNA sequence at one arm of the stem and a respective antisense sequence at the opposite arm. Finally, precursor candidate sequences were checked using the BLASTn algorithm from the miRBase (www.miRBase.org) and NCBI databases.

For the frequency analysis of all identified miRNAs, sRNA reads were aligned in Bowtie v 0.12.7 using the default parameters, with the first seed alignment >28 nt in size and allowing zero mismatches. As reference, we used both previously annotated pre-miRNAs from miRBase and the putative pre-miRNAs identified in this work. The SAM files from Bowtie were then processed using Python scripts to assign the frequencies of each read and map them onto references. For data normalization, we use the scaling normalization method proposed by [49]. To assess whether the microRNA was differentially expressed, we independently used both the R package EdgeR [50] and the A–C test [51]. In brief, EdgeR uses a negative binomial model to estimate overdispersion from the miRNA count. The dispersion parameter of each

miRNA was estimated by the tagwise dispersion. Then, differential expression is assessed for each miRNA using an adapted exact test for overdispersed data. The A-C test computes the probability that two independent counts of the same microRNA came from similar samples. We considered miRNAs to be differentially expressed if they had a p-value ≤ 0.001 in both statistical tests.

Prediction of miRNA Targets

The prediction of target genes of novel miRNAs was performed against assembled RNA-seq contigs using psRNAtarget [52], with the default parameters and a maximum expectation value of 4 (number of mismatches allowed). Candidate RNA sequences were then annotated by assigning them putative gene descriptions based on their sequence similarity with previously identified and annotated genes that had been deposited in the NR and Swiss-Prot/Uniprot protein databases using BLASTx; this analysis was conducted using the blast2GO v2.3.5 software [53]. The annotation was improved by analyzing conserved domains/families using the InterProScan tool. Gene Ontology (GO) terms for the cellular component, molecular function and biological processes were determined using the GOslim tool in the blast2GO software. The orientation of the transcripts was obtained from BLAST annotations.

Results

Overview of B. napus RNAs Library Sequencing

To identify the miRNA transcriptome involved in *B. napus* seed development, sRNA libraries constructed from mature seeds and from an equivalent mix of immature seeds (a pool of DAF stages 7–42) were sequenced by using the Solexa/Illumina platform. Deep sequencing yielded a total of almost 38 million sRNA reads. After removing low-quality sequences, adapter contaminants and inserts, approximately 17 million and 19 million reads, with lengths ranging from 18 to 25 nt, were obtained from the mature and developing seed libraries, respectively (Table 1); these reads represented 8,632,807 and 5,665,721 of distinct sequences in each library, respectively (Table S1). Consistent with the length distribution pattern of sRNAs in other plant species, sequences between 21 to 24 nt long were the most abundant, with 24 nt long sRNAs as the main peak (Figure 1). A relatively large number of 22 and 23 nt long small RNAs were obtained in the developing seed dataset. This was previously observed in developing *B. napus* seed sRNA libraries [41]. The highest sequence redundancy was observed in the 21 nt long fraction of mature seed library and the 24 nt long fraction of the developing seed library (Figure 1 and Table S1). A small fraction from the total number of reads sequenced in the mature and developing seed libraries (10.2% and 2.2%, respectively) matched to miRNAs (Table 2). Approximately 4.3% and 2.9% of the reads matched non-coding sRNAs other than miRNAs (rRNA, tRNA, snRNA, snoRNA), respectively, and 3.7% and 0.5% matched organellar sRNAs (mtRNA, cpRNA), respectively. The majority of the reads did not match known small RNAs and possibly represent siRNAs.

Because the genome of *B. napus* is not publicly available, we sequenced the mRNA transcriptome of *B. napus* seeds for use as a reference sequence in further analysis. The pooled mRNA-seq yielded 32,485,023 reads, which were imported into the CLC Genomics Workbench and *de novo* assembled into 237,993 contigs with an average length of 284 bp. Contigs and non-assembled reads with a minimum length of 100 bp were further considered. The contigs ranged in size between the minimum set threshold of 100 bp and 12,344 bp (average size = 285 bp; N50 = 361 bp), with 29,157 contigs that were more than 500 bp in length.

Identification of Conserved miRNAs in B. napus Seeds

There were 4,680 mature miRNAs from 52 Viridiplantae species deposited in the miRBase Release 18.0 from November 2011 [48]. Because miRNAs are highly conserved among plant species, the first approach to characterize the miRNA libraries was to precisely identify miRNAs by sequence homology. To identify conserved miRNAs in *B. napus* (Bna), the libraries were matched against the complete set of 2,585 unique plant mature miRNAs sequences from miRBase with no mismatches allowed. In total, 1,949,940 reads perfectly matched 219 known mature miRNA sequences, which corresponded to 45 plant miRNA families. On average, four miRNA members were identified within each miRNA family (Figure 2). Mature sequences matching MIR156 and MIR57, MIR165 and MIR166 or MIR170 and MIR171 were grouped as one single family due to their shared evolutionary origin. Of these reads, a total of 196 miRNAs were identified in the mature seed library, and 172 miRNAs were identified in the developing seed library, while 149 miRNAs were shared by both libraries (Table S2). From the total of 48 mature miRNAs annotated in miRBase for *B. napus* (Bna-miRNA), 24 unique Bna-miRNAs were detected in the libraries, representing all 17 known Bna-miRNA families. The remaining 28 miRNA families comprised miRNAs that are newly identified in *B. napus* but conserved in Brassicaceae species or among several plant species (Table 3 and Table S2). Overall, the largest family was MIR156/157, with 24 members representing MIR156/157 variants found in different species. MIR165/166 (21 members), MIR169 (15 members) and MIR319 (14 members) were the second, third and fourth largest miRNA families, respectively. Of the remaining miRNA families, 19 contained between 2 to 6 members, while 17 were represented by a single member.

Identification of Novel B. napus miRNAs

To distinguish miRNAs from other small RNAs, such as siRNAs, some important features from miRNA biogenesis must be considered: 1) mature miRNAs are derived from pre-miRNAs; 2) all pre-miRNAs can form a secondary structure with a stem-looped hairpin; 3) the secondary structure shows high negative minimum folding free energy (MFE, 40–100 kcal/mol) and minimum folding free energy index (MFEI, higher than 0,85) [54]; 4) The stem-looped hairpin has the mature miRNA sitting on one of the arms and an almost complementary miRNA (with few mismatches) on their opposite site arm (5p and 3p positions). To identify novel miRNAs in *B. napus* seeds, the sRNA libraries were matched against assembled contigs of developing and mature seeds, because Bna ESTs and GSS were previously explored elsewhere [25,24,29,41]. Candidate mRNA sequences with hairpin-like structures and with more than 10 miRNA reads that were anchored in the same orientation in the 5p and/or 3p arm in a two block-like pattern were considered putative pre-miRNAs. The MFE and MFEI were determined for each candidate sequence and the precursor identity was determined by BLAST searches against mature miRNAs at miRBase. As a result, three groups of pre-miRNAs were identified: (a) known in Bna, (b) new in Bna but known in plants and (c) new in Bna and uncharacterized in other plants.

The determined secondary structures of Bna pre-miRNAs identified in the first group showed an average MFE value of -57.16 kcal/mol, an average MFEI of -0.99 and an average GC content of 43.32% (Table S3). In addition, four mRNA sequences presented anchored miRNAs in a block-like pattern (Bna-MIR393-2; Bna-MIR393-2; Bna-MIR396; Bna-MIR1140) but did not fold into a secondary structure because they had partial mRNA sequences. However, these four sequences were considered

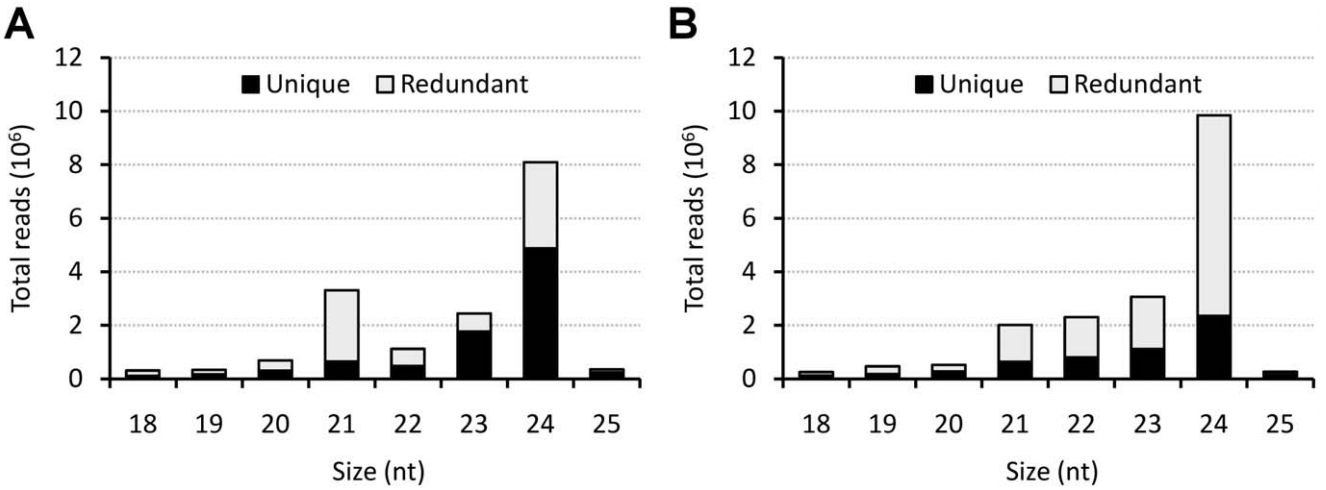

Figure 1. Length distribution and diversity of total number of small RNA reads of *B. napus* seed libraries. (**A**) Small RNA reads sequenced from the mature seed library. (**B**) Small RNA reads sequenced from the developing seed library.
doi:10.1371/journal.pone.0050663.g001

an exception and were studied further because they showed high similarity to known Bna pre-miRNAs. In total, 17 new full-length and 4 partial pre-miRNA sequences were identified for 11 known Bna-miRNA families, along with 16 new 5p:3p pairs of mature Bna-miRNAs (Table 3 and Figure S2). It has been previously shown that miRNA variants, referred to as isomiRNAs, are detectable using high-throughput sequencing [36,38,55]. Known Bna-miRNAs and several novel isomiRNAs were detected in the predicted precursors (Table S3). The known Bna sequences represented the most abundant miRNA in eight of the 21 new precursors (Bna-MIR159, Bna-MIR166, Bna-MIR167, Bna-MIR168, Bna-MIR171 and Bna-MIR824) and were not considered new miRNAs in Table 3. The second group of new pre-miRNAs (Bna-nMIRx) comprised seven full-length and one partial pre-miRNA. Mature miRNA sequences of seven miRNA families (Bna-nMIR158, Bna-nMIR162, Bna-nMIR172, Bna-nMIR394, Bna-nMIR400, Bna-nMIR408 and Bna-nMIR827) that were not previously characterized in *B. napus* have been identified in these new pre-miRNAs (Table 3 and Table S4). The miRNA families MIR158 and MIR400 have been reported only in Brassicaceae species, whereas the other families are conserved in several plants. With the exception of one partial pre-miRNA (Bna-nMIR394), all of the new pre-miRNAs had 5p:3p arm miRNA pairs that were complementarily anchored (Figure S3). Several isomiRNAs were detected and are shown in Figure S3. Bna-nMIR827 showed one mismatch with other plant miRNAs and therefore it has not been detected on initial analysis (Table S2). The third group of pre-miRNAs comprised all sequences with characteristic hairpin-like structures with no homology to previously known plant miRNAs; these sequences were considered as novel pre-miRNAs in plants. To increase the reliability of the predictions, one additional criterion was considered: only candidate precursors with anchored mature sequences that could be found in both libraries or for which a complementary miRNA sequence could be identified in at least one library were annotated. As a result, 15 novel miRNAs, representing 13 novel Bna-miRNA families and distributed in 15 new precursors, were identified (Table 3). From these new miRNAs, 11 pre-miRNAs exhibited the 5p:3p miRNA pair (Figure S3). The average MFE value of the 23 newly predicted pre-miRNAs (plant conserved and novel) was −46.67 kcal/mol with a range of −8.7 to −131.5 kcal/mol (Table S4). The average length of the pre-miRNAs was 131 nt with a MFEI of −0.92 and a GC content of 39%. In accordance with previous results [33], the majority of the newly identified miRNA sequences had uracil (U) as their first nucleotide (Table S4).

Expression Profile of B. napus miRNAs

The large number of sequences produced by high-throughput sequencing enables the use of read counts in libraries as a reliable source for estimating the abundance of miRNAs [56,57,58]. The most abundant miRNAs identified by sequence homology in the libraries were MIR156, MIR159, MIR166, MIR167 and MIR824, each with more than 100,000 reads sequenced (Figure 3a). The majority of the conserved miRNAs that were

Table 1. Summary of sequencing data of *B. napus* small RNA libraries.

	Mature seeds		Developing seeds	
Reads	**Number of reads**	**Percentage (%)**	**Number of reads**	**Percentage (%)**
Total reads*	17,878,538	100.0	19,954,089	100.0
18–25 nt	16,658,523	93.2	18,728,461	93.9
<18 nt	875,194	4.9	856,483	4.3
>25 nt	344,821	1.9	369,145	1.8

*High quality reads with lengths of 1 to 44 nt.
doi:10.1371/journal.pone.0050663.t001

Table 2. Categorization of *B. napus* sequecences matching noncoding and organellar small RNAs.

	Mature seeds		Developing seeds	
sRNA*	**Number of reads**	**Percentage (%)**	**Number of reads**	**Percentage (%)**
miRNA	1,699,293	10.20	420,230	2.24
rRNA	675,151	4.05	524,132	2.80
tRNA	39,769	0.24	23,449	0.13
snRNA	2,688	0.02	1,830	0.01
snoRNA	1,911	0.01	1,567	0.01
mtRNA	298,127	1.79	44,370	0.24
cpRNA	316,543	1.90	51,188	0.27
other sRNA	13,625,041	81.79	17,661,695	94.30
Total	16,658,523	100	18,728,461	100

*Only 18–25 nt reads were considered. The small RNA were clustered according to their origin as follow: ribosome (rRNA); transporter (tRNA); small nuclear (snRNA); small nucleolar (snoRNA); mitochondrial (mtRNA) and chloroplastic (cpRNA).

doi:10.1371/journal.pone.0050663.t002

identified had been sequenced less than 1,000 times, and 11 miRNA families had been detected less than 10 times. Although the total number of unique miRNAs detected in both libraries were similar, the number of total reads was higher in the mature seed library, where 1,581,402 reads (196 miRNAs) were identified, compared to 368,538 reads (172 miRNAs) in the developing seed library. A few poorly represented miRNA families, namely, MIR828 and MIR2111, were predominantly detected in the developing seed library (Figure 3a). Sharp differences in read abundance were also observed within members of one family and between miRNA libraries. For example, the abundance of MIR156/157 members ranged from 2 to 155 844 reads in the mature library and from 1 to 2 135 in the developing library. Comparisons between the normalized data suggested that 79 conserved miRNAs were differentially represented between the two libraries (Table S2). Differentially represented miRNAs in the developing seed library that exhibited more than a 2-fold change are shown in Figure 3b. Some members of MIR156/157, MIR162, MIR164, MIR168, MIR169, MIR172, MIR393, MIR395, MIR396, MIR398, MIR399, MIR828 and MIR1140 were more represented in developing seeds than in mature seeds (Figure 3b). On the contrary, some members of MIR156/157, MIR169, MIR319, MIR390, MIR391, MIR403, MIR824 and MIR1885 were more represented in mature seeds than in developing seeds (Figure 3b).

Target Prediction of B. napus miRNAs

To infer the biological functions of the 23 newly identified miRNAs (plant conserved and novel), putative target genes were searched. The most abundant mature miRNAs were aligned to assembled *B. napus* contigs using the web-based computer server psRNATarget. Default parameters and a maximum expectation value of 4 (number of mismatches allowed) were used for higher prediction coverage. A total of 105 contigs matched miRNAs of the 14 novel and 8 known plant miRNA families identified in *B. napus*, representing 89 unique potential targets with an average of four targets per miRNA molecule. All of the identified targets were analyzed using a BLASTX against protein databases, followed by GO analysis to evaluate their putative functions. The detailed results of the best BLASTX hits are shown in Table S5. According to the categorization of GO annotation, 103 genes are involved in cellular components, with the majority of them localized in the nucleus and organelles. In the category of molecular functions, 103 genes participate in catalytic activities and binding activities with proteins and nucleic acids. With respect to biological processes, 95 genes primarily took part in responses to stimulus and different cellular and metabolic processes, suggesting that the novel Bna-miRNAs are involved in a broad range of physiological functions (Figure S4). We searched the putative target genes for differentially represented miRNAs and isomiRNAs shown in Table S2 and S3 to investigate whether these miRNAs regulate target genes

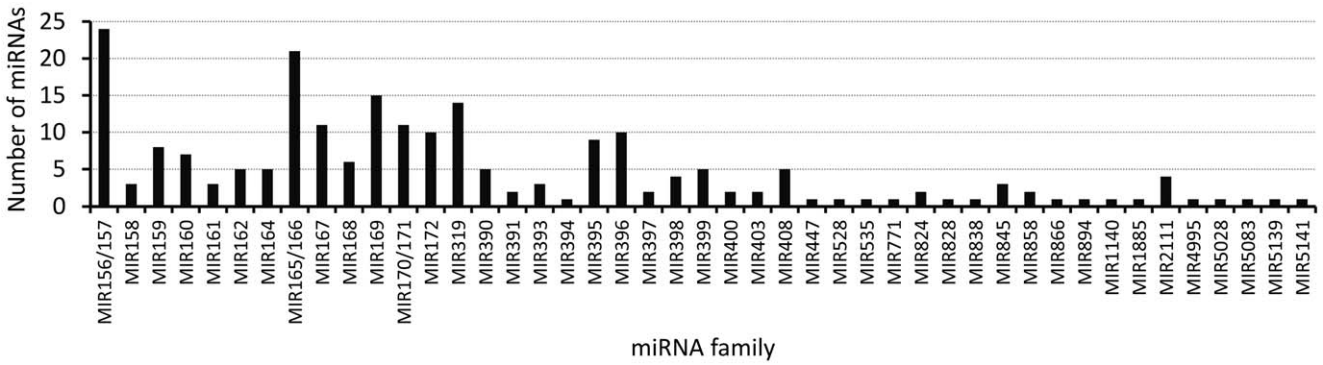

Figure 2. Number of miRNAs identified in *B. napus* seed libraries in known miRNA family in plants. The numbers are the sum of different miRNA containing the canonical sequences from the families of plant miRNAs deposited in miRBase.

doi:10.1371/journal.pone.0050663.g002

Table 3. Number of miRNAs identified by sequence homology or matching pre-miRNAs in *B. napus* seed libraries that belong to novel and known plant miRNA families.

	Size									
Class	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**Total**	**Precursors**	**Families**
New miRNAs known in other plants species (without precursor)[a]	0	7	40	122	17	0	2	**188**	0	**28**
Known miRNAs in B. napus (without precursor)[a]	0	0	0	21	3	0	0	**24**	0	**17**
New miRNAs in known *B. napus* families* (with precursor)[b]	0	1	7	7	1	0	0	**16**	21	**0 (11)[d]**
New miRNAs known in other plants species* (with precursor)[c]	0	0	1	6	1	0	0	**8**	8	**1 (6)[d]**
New miRNAs unknown in other plants species* (with precursor)[c]	1	0	0	12	2	0	0	**15**	15	**13**
Total	**1**	**8**	**49**	**171**	**24**	**0**	**2**	**251**	**44**	**59**

*most abundant;
[a]Data from Table S2;
[b]Data from Table S3;
[c]Data from Table S4;
[d]Number of identified families already considered in previous categories are in parenthesis.
doi:10.1371/journal.pone.0050663.t003

involved in seed development and energy storage in *B. napus*. We found that 313 contigs were potential targets of 44 overrepresented miRNAs and 221 contigs were potential targets of 36 underrepresented miRNAs in the developing library (Table S6). In total, an average of seven targets per miRNA molecule was identified. According to the categorization of GO annotation, 436 genes are involved in cellular components, and 489 genes have been classified within categories of molecular function (Figure 4). With respect to biological processes (424 genes), miRNAs that were more abundant in mature than in developing seeds were found to potentially target genes that took part in growth, developmental processes, multicellular organismal process and biological regulation, along with different cellular and metabolic processes and responses to stimulus (Figure 4).

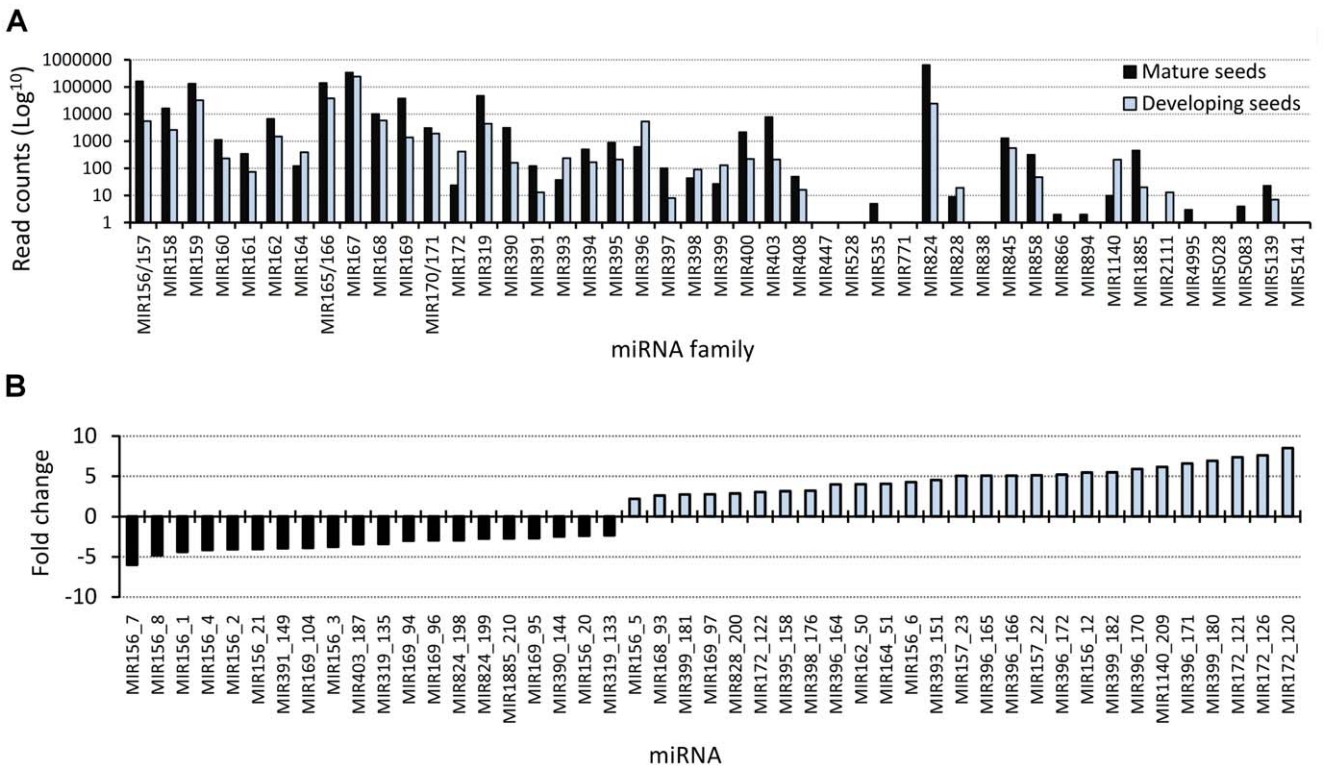

Figure 3. Sequencing profile of plant conserved miRNAs in *B. napus* seed libraries. (**A**) Number of total read counts of each miRNA in the mature and developing seed libraries of *B. napus*. (**B**) Mature miRNAs differentially expressed in the developing seed library and with fold-change higher than 2,0. Black bars represent miRNAs that were more abundant in mature seeds; blue bars represent miRNAs that were more abundant in developing seeds.
doi:10.1371/journal.pone.0050663.g003

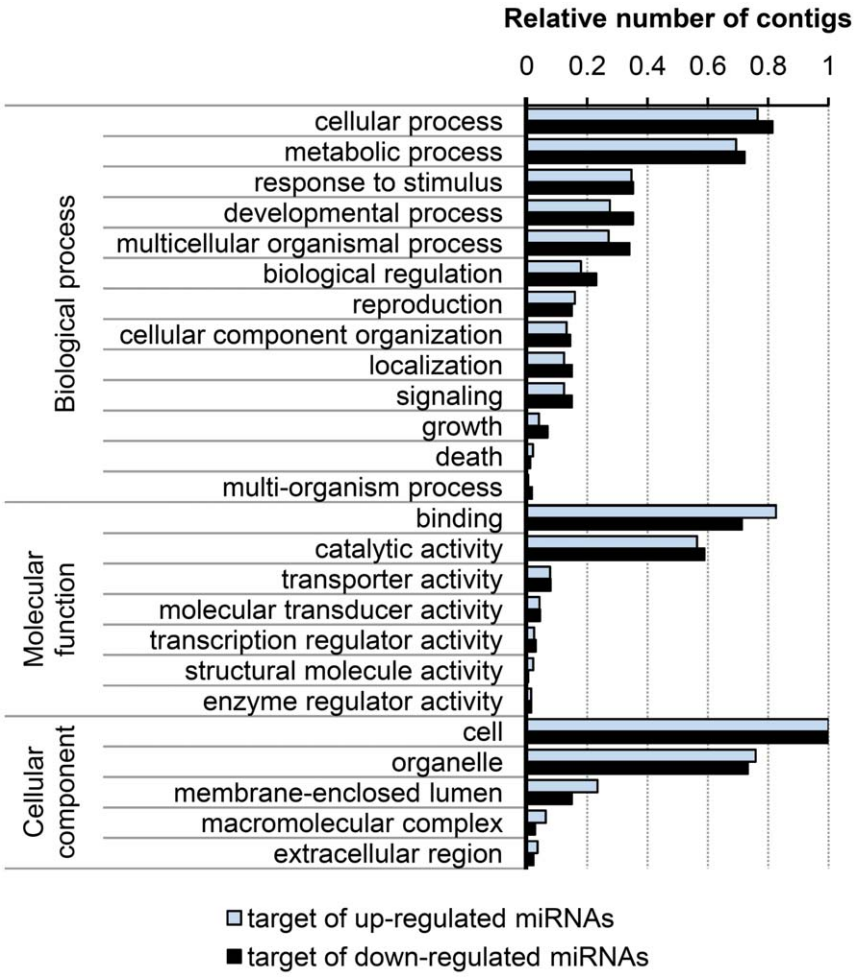

Figure 4. Targets of differentially expressed miRNAs in developing seeds of *B. napus*. The total number of contigs for each Gene Ontology (GO) term is relative to the total number of contigs in each gene category.
doi:10.1371/journal.pone.0050663.g004

Discussion

MiRNAs have been shown to play critical roles in the regulation of gene expression during plant development and in species-specific adaptation processes. In this study, we profiled by deep sequencing the microRNAome of the mature and immature stages of *B. napus* seeds. A total of 59 miRNA families were detected in the sRNA libraries (Tables S2 and 3). The families detected here increase the number of 17 miRNA families previously described in *B. napus* in the miRBase registry. We describe 29 miRNA new families in *B. napus* but conserved in other plants, and 13 families that were reported for the first time in plants.

A large number of reads were sequenced from both miRNA libraries of developing seeds and mature seeds of *B. napus*, providing a good representation of the miRNA population in seeds (Table 1). As expected, most of the highly conserved miRNAs in diverse plant species were also the most abundant in *B. napus* seeds. In addition, the conserved miRNA families of *B. napus* showed the higher number of members per family [30,31]. miRNA families described only in Brassicaceae species were identified (MIR158, MIR161, MIR391, MIR400, MIR447, MIR771, MIR824, MIR838, MIR858 and MIR866) along with two families (MIR1885 and MIR1140) that could be specific to the genus *Brassica*. MIR1885, which was previously only identified in *B. rapa*, has been detected in the present *B. napus* libraries, and Bna MIR1140 was recently detected in *B. rapa* [39]. These results suggest that both the ancient regulatory pathways mediated by evolutionarily conserved miRNAs as well as novel and specialized pathways unique to Brassicaceae species, are present in *B. napus* [30,31]. Nearly all of the unique Bna-miRNA sequences described in miRBase were detected, and all of the Bna-miRNA families were represented in at least one library. In addition, the identification of Bna pre-miRNAs allowed for the identification of several isomiRNAs that were identical to some conserved miRNAs identified in Table S2 and often more abundant than the known Bna sequences (Tables S2 and S3). Because the Bna-miRNA sequences deposited in miRBase were mainly identified by sequence homology and cloning of miRNAs isolated from whole plant tissues, it is tempting to conclude that the most abundant miRNA sequences detected in this study are seed specific. Although similar conclusions were proposed in rice [34], this observation is likely to reflect the increased detection power of the deep sequencing strategy and the limited computational analysis due to an incomplete genome.

To predict new miRNAs with confidence, we identified precursor sequences according to the strict criteria of having sharply defined distribution patterns of one or two block-like anchored sRNAs and at least 11 reads in total. It was previously

found that the read depth distribution along putative pre-miRNAs can be used as a reliable guide for differentiating possible miRNAs from contaminant sequences, such as degradation products of mRNAs or transcripts that are simultaneously expressed in both the sense and antisense orientations [59]. In addition, the average MFE and MFEI of the predicted stem-loop structures of the pre-miRNAs values were within the confidence values suggested by [54] and are similar to the length average, MFE, MFEI and GC content of pre-miRNAs from other plant species, such as Arabidopsis [60]. For the majority of the new miRNAs, the complementary miRNA species (5p:3p pairs) were detected in our libraries, providing strong evidence that they derived from precisely processed stem-loops during miRNA biogenesis [6]. During the preparation of this manuscript, [41] reported nine new miRNA families in the very early stages of *B. napus* seed development. Bna-nMIR03, which was detected in both the developing and mature libraries in the present study, showed an identical sequence to one of the miRNA families presented by [41]. Taken together, these results strongly suggest that genuinely new Bna-miRNAs have been predicted, and also demonstrate that using a combination of sRNA and mRNA sequencing is a powerful strategy to discover new miRNAs in plants without an available genome.

In this study, 23 new miRNAs have been identified in *B. napus* seed libraries (Table S4). Furthermore, several miRNAs and isomiRNAs were more represented in developing seeds and may regulate the expression of target genes involved in seed development and energy storage in *B. napus* (Figure 3b, Tables S2 and S3). To infer about the biological significance of the results, *in silico* target predictions with the permissive expectation value of 4 were chosen. This strategy, which can include false targets, were previously used to successful predict true alternative targets that can be species or tissue-specific [61,62]. GO annotation analyses suggested that miRNAs more abundantly present in mature seeds are probably involved in the down-regulation of genes that are more important during seed development, namely genes related to auxin signaling (ARFs, F-boxes, auxin efflux carrier component) or essential transcription factors in the regulation of plant development (NAC, SCL, TOE) [63–65]. Because the accumulation of dry matter for seed germination is the main priority of developing seeds, a large number of target genes may participate in these processes. Interestingly, some of the targets from the differentially abundant MIR156, MIR167, MIR169, MIR171, MIR319 and MIR396 were related to lipid metabolism (Table S6). Defects in ethanolamine-phosphate cytidylyltransferase, which is the target of Bna-nMIR04 and is predicted to be involved in lipid metabolic process, have been shown to affect embryonic and postembryonic development in Arabidopsis [66]. In conclusion, some potentially valuable targets emerge from the analysis that would be interesting to validate. Further investigation in the role of seed-specific miRNAs will contribute to the knowledge of the energy storage process in seeds.

This work provides a comparative study of the miRNA content of the transcriptome of mature and developing seeds of *B. napus*. The results will support future research on deeper studies of individual Bna-miRNAs and their function on embryogenesis and seed maturation. It is clear that the identification of miRNAs is not yet complete in *B. napus* and that the release of the genome sequences will be essential to fully understand the complete miRNA repertoire. One future endeavor is to look for more novel miRNAs; however, expression analysis and target validation will be critical for determining the biological functions of both the conserved and novel miRNAs identified during each developing seed stage of different *B. napus* cultivars.

Accession Numbers

Sequencing data is available at the NCBI Gene Expression Omnibus (GEO) ([http://www.ncbi.nlm.nih.gov/geo]). The accession number GSE38020 contain the sequence data of mature and developing seed libraries from mRNA-seq and sRNA-seq.

Supporting Information

Figure S1 Flow chart of the procedure for the identification of miRNAs.
(TIF)

Figure S2 Predicted secondary structures of the new pre-miRNAs of known *B. napus* miRNA families. Secondary structures and the locations of the miRNAs mapped onto these precursors. Mature miRNAs located in the 5p and 3p arms are labeled in magenta and red, respectively.
(PDF)

Figure S3 Predicted secondary structures of new pre-miRNAs in *B. napus*. Secondary structures of candidate miRNA precursors of new *B. napus* miRNA families (Bna-nMIRx), their locations and the abundance of small RNAs mapped onto these precursors. Mature miRNAs located in the 5p and 3p arms are labeled in magenta and red, respectively. Values on the left side of the miRNA sequence represent read counts in the mature and developing seed libraries, respectively.
(PDF)

Figure S4 Distribution of target genes of new Bna-miRNAs in gene categories and Gene Ontology (GO) terms.
(TIF)

Table S1 Length distribution of raw reads of *B. napus* small RNAs libraries.
(XLS)

Table S2 New and Bna-known miRNAs identified in small RNA libraries. *B. napus* miRNAs with perfect homology to plant miRNAs deposited in the miRBase database (release 18, November 2011).
(XLS)

Table S3 New and Bna-known putative miRNA precursors identified in *B. napus* miRNA families.
(XLS)

Table S4 New miRNAs and pre-miRNAs identified in *B. napus* libraries. Sequences were classified as known or as novel plant miRNA families.
(XLS)

Table S5 Predicted targets of novel and plant conserved miRNAs in *B. napus*.
(XLS)

Table S6 Predicted targets of differentially expressed miRNAs in *B. napus*.
(XLS)

Author Contributions

Conceived and designed the experiments: RM MMP. Performed the experiments: APK FG LFVO FSM. Wrote the paper: APK. Performed some of the data analysis: MPA RDM GLM. Revised the paper: ACTZ AC FSM.

References

1. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116: 281–297.
2. Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9: 102–114.
3. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 5: 522–531.
4. Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. Cell 136: 669–687.
5. Lee Y, Kim M, Han J, Yeom KH, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23: 4051–60.
6. Kurihara Y, Watanabe Y (2004) Arabidopsis microRNA biogenesis through Dicer-like 1 protein functions. Proc Natl Acad Sci USA 101: 12753–12758.
7. Mallory AC, Bartel DP, Bartel B (2005) MicroRNA-directed regulation of Arabidopsis AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes. Plant Cell 17: 1360–1375.
8. Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAs and their regulatory roles in plants. Ann Rev Plant Biol 57: 19–53.
9. Mallory AC, Vaucheret H (2006) Functions of microRNAs and related small RNAs in plants. Nat Genet 38: S31–6.
10. Poethig RS (2009) Small RNAs and developmental timing in plants. Curr Opin Genet Dev 19: 374–378.
11. Rubio-Somoza I, Weigel D (2011) MicroRNA networks and developmental plasticity in plants. Trends Plant Sci 16: 258–64.
12. Nodine M, Bartel DP (2010) MicroRNAs prevent precocious gene expression and enable pattern formulation during plant embryogenesis. Genes Dev 24: 2678–2692.
13. Willmann MR, Mehalick AJ, Packer RL, Jenik PD (2011) microRNAs regulate the timing of embryo maturation in Arabidopsis. Plant Physiology 155: 1871–1884.
14. Chen X (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science 303: 2022–2025.
15. Lauter N, Kampani A, Carlson S, Goebel M, Moose SP (2005) microRNA172 downregulates glossy15 to promote vegetative phase change in maize. Proc Natl Acad Sci USA 102: 9412–9417.
16. Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science 297: 2053–2056.
17. Bartel DP (2009) MicroRNAs: Target recognition and regulatory functions. Cell 136: 215–233.
18. Huntzinger E, Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nature Reviews Genetics 12: 99–110.
19. Appelqvist LA (1972) Chemical composition of rapeseed. In: Appelqvist LA, Ohlson R, editors. Rapeseed. Amsterdam: Elsevier. 123–173.
20. Norton G, Harris JF (1983) Triacylglycerols in oilseed rape during seed development. Phytochemistry 22: 2703–2707.
21. He Y-Q, Wu Y (2009) Oil Body Biogenesis during Brassica napus Embryogenesis. J of Integrative Plant Biol 51: 792–799.
22. Troncoso-Ponce MA, Kilaru A, Cao X, Durrett TP, Fan J, et al. (2011) Comparative deep transcriptional profiling of four developing oilseeds. Plant J 68: 1014–1027.
23. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36: D154–D158.
24. Xie FL, Huang SQ, Guo K, Xiang AL, Zhu YY, et al. (2007) Computational identification of novel microRNAs and targets in *Brassica napus*. FEBS Lett 581: 1464–74.
25. Wang L, Wang MB, Tu JX, Helliwell CA, Waterhouse PM, et al. (2007) Cloning and characterization of microRNAs from *Brassica napus*. FEBS Lett 581: 3848–3856.
26. Buhtz A, Springer F, Chappell L, Baulcombe DC, Kehr J (2008) Identification and characterization of small RNAs from the phloem of *Brassica napus*. Plant J 53: 739–749.
27. Pant BD, Musialak-Lange M, Nuc P, May P, Buhtz A, et al. (2009) Identification of nutrient-responsive *Arabidopsis* and rapeseed microRNAs by comprehensive real-time polymerase chain reaction profiling and small RNA sequencing. Plant Physiol 150: 1541–1555.
28. Buhtz A, Pieritz J, Springer F, Kehr J (2010) Phloem small RNAs, nutrient stress responses, and systemic mobility. BMC Plant Biol 13: 10: 64.
29. Huang SQ, Xiang AL, Che LL, Chen S, Li H, et al. (2010) A set of miRNAs from *Brassica napus* in response to sulphate deficiency and cadmium stress. Plant Biotechnol J 8: 887–899.
30. Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, et al. (2010) MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. Plant Cell 22: 1074–89.
31. Lenz D, May P, Walther D (2011) Comparative analysis of miRNAs and their targets across four plant species. BMC Res Notes 8: 4: 483.
32. Moxon S, Jing R, Szittya G, Schwach F, Rusholme Pilcher RL, et al. (2008) Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. Genome Res 18: 1602–1609.
33. Subramanian S, Fu Y, Sunkar R, Barbazuk WB, Zhu JK, et al. (2008) Novel and nodulation-regulated microRNAs in soybean roots. BMC Genomics 9: 160.
34. Zhu QH, Spriggs A, Matthew L, Fan L, Kennedy G, et al. (2008) A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Res 18: 1456–1465.
35. Hsieh LC, Lin SI, Shih AC, Chen JW, Lin WY, et al. (2009) Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. Plant Physiol 151: 2120–2132.
36. Lelandais-Brière C, Naya L, Sallet E, Calenge F, Frugier F, et al. (2009) Genome-wide *Medicago truncatula* small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules. Plant Cell 21: 2780–2796.
37. Li Y, Zhang Q, Zhang J, Wu L, Qi Y, et al. (2010) Identification of microRNAs involved in pathogen-associated molecular pattern-triggered plant innate immunity. Plant Physiol 152: 2222–2231.
38. Kulcheski FR, de Oliveira LF, Molina LG, Almerão MP, Rodrigues FA, et al. (2011) Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC Genomics 10: 12: 307.
39. Yu X, Wang H, Lu Y, de Ruiter M, Cariaso M et al. (2012) Identification of conserved and novel microRNAs that are responsive to heat stress in *Brassica rapa*. J Exp Bot 63: 1025–38.
40. De Paola D, Cattonaro F, Pignone D, Sonnante G (2012) The miRNAome of globe artichoke: conserved and novel micro RNAs and target analysis. BMC Genomics 24: 13: 41.
41. Zhao YT, Wang M, Fu SX, Yang WC, Qi CK, et al. (2012) Small RNA profiling in two *Brassica napus* cultivars identifies microRNAs with oil production- and development-correlated expression and new small RNA classes. Plant Physiol 158: 813–23.
42. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864.
43. Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF et al. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res 37: D159–D162.
44. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196.
45. He S, Liu C, Skogerbø G, Zhao H, Wang J, et al. (2008) NONCODE v2.0: decoding the non-coding. Nucleic Acids Res 36: D170–D172.
46. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
47. Li, R. Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–7.
48. Moxon S, Schwach F, Maclean D, Dalmay T, Studholme DJ, et al. (2008) A tool kit for analysing large-scale plant small RNA datasets. Bioinformatics 24: 2252–2253.
49. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25.
50. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–40.
51. Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res 7: 986–995.
52. Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 39: W155–9.
53. Conesa A and Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008: 1–13.
54. Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA (2006) Evidence that miRNAs are different from other RNAs. Cell Mol Life Sci 63: 246–54.
55. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, et al. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18: 610–621.
56. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, et al. (2007) High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. PLoS ONE 2: e219.
57. Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, et al. (2009) Limitations and possibilities of small RNA digital gene expression profiling. Nat Methods 6: 474–6.
58. McCormick KP, Willmann MR, Meyers BC (2011) Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. Silence 2: 2.
59. Schreiber AW, Shi BJ, Huang CY, Langridge P, Baumann U (2011) Discovery of barley miRNAs through deep sequencing of short reads. BMC Genomics 25: 12: 129.
60. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev 16: 1616–1626.
61. Jones-Rhoades MW, Bartel DP (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Molecular Cell 14: 787–799.

62. Debernardi JM, Rodriguez RE, Mecchia MA, Palatnik JF (2012) Functional specialization of the plant miR396 regulatory network through distinct microRNA-target interactions. PLoS Genetics 8: e1002419.
63. Mallory AC, Dugas DV, Bartel DP, Bartel B (2004) MicroRNA regulation of NAC-domain targets is required for proper formation and separation of adjacent embryonic, vegetative, and floral organs. Curr Biol 14: 1035–1046.
64. Zhang ZL, Ogawa M, Fleet CM, Zentella R, Hu J, et al. (2011) Scarecrow-like 3 promotes gibberellin signaling by antagonizing master growth repressor DELLA in Arabidopsis. Proc Natl Acad Sci U S A 108: 2160–5.
65. Huijser P, Schmid M (2011) The control of developmental phase transitions in plants. Development 138: 4117–29.
66. Mizoi J, Nakamura M, Nishida I (2006) Defects in CTP:phosphorylethanolamine cytidylyltransferase affect embryonic and postembryonic development in Arabidopsis. Plant Cell 18: 3370–85.

Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps

Sanjeev Kumar Sharma,*[,1] Daniel Bolser,[†,1,2] Jan de Boer,[‡] Mads Sønderkær,[§] Walter Amoros, Martin Federico Carboni,[††] Juan Martín D'Ambrosio,[‡‡] German de la Cruz,[‡‡] Alex Di Genova,[§§] David S. Douches,*** Maria Eguiluz,[†††] Xiao Guo,[‡‡‡] Frank Guzman,[†††,3] Christine A. Hackett,[§§§] John P. Hamilton,**** Guangcun Li,[‡‡‡] Ying Li,[††††] Roberto Lozano,[†††] Alejandro Maass,[§§] David Marshall,[‡‡‡‡] Diana Martinez,[†††] Karen McLean,* Nilo Mejía,[§§§§] Linda Milne,[‡‡‡‡] Susan Munive,** Istvan Nagy,*****[,4] Olga Ponce,[†††] Manuel Ramirez,[†††] Reinhard Simon,** Susan J. Thomson,[††††] Yerisf Torres,[†††] Robbie Waugh,* Zhonghua Zhang,[††††] Sanwen Huang,[††††] Richard G. F. Visser,[‡] Christian W. B. Bachem,[‡] Boris Sagredo,[‡‡‡‡‡] Sergio E. Feingold,[††] Gisella Orjeda,[†††] Richard E. Veilleux,[§§§§§] Merideth Bonierbale,** Jeanne M. E. Jacobs,[†††††] Dan Milbourne,***** David Michael Alan Martin,[†] and Glenn J. Bryan*[,5]**

*Cell and Molecular Sciences and [‡‡‡‡]Information and Computational Sciences, The James Hutton Institute, Dundee DD2 5DA, United Kingdom, [†]Division of Biological Chemistry and Drug Discovery, College of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom, [‡]Laboratory of Plant Breeding, Department of Plant Sciences, Wageningen-UR, 6708 PB Wageningen, The Netherlands, [§]Department of Biotechnology, Chemistry and Environmental Engineering, 9000 Aalborg University, Aalborg, Denmark, **International Potato Center (CIP), Lima 12, Peru, [††]Laboratorio de Agrobiotecnología, Instituto Nacional de Tecnología Agropecuaria (INTA) cc276 (7620) Balcarce, Argentina, [‡‡]Laboratorio de Genética y Biotecnología Vegetal, Universidad Nacional San Cristobal de Huamanga, Ayacucho 05000, Perú, [§§]Mathomics, Centro de Regulación Genómica & Centro de Modelamiento Matemático, Universidad de Chile, Santiago 8320000, Chile, ***Department of Crop and Soil Sciences and ****Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824, [†††]Genomics Research Unit, Facultad de Ciencias, Universidad Peruana Cayetano Heredia, Lima 31, Peru, [‡‡‡]Institute of Vegetables, Shandong Academy of Agricultural Sciences, Jinan 250100, China, [§§§]Biomathematics and Statistics Scotland, Dundee DD2 5DA, United Kingdom, [††††]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China, [§§§§]INIA-La Platina, Santiago 8831314, Chile, *****Crops Environment and Land Use Programme, Teagasc, Carlow, Ireland, [†††††]The New Zealand Institute for Plant & Food Research Ltd., Christchurch 8120, New Zealand, [‡‡‡‡‡]INIA-Rayentué, Rengo 2940000; Universidad de la Frontera, Temuco 4811230, Chile, and [§§§§§]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

ABSTRACT The genome of potato, a major global food crop, was recently sequenced. The work presented here details the integration of the potato reference genome (DM) with a new sequence-tagged site marker–based linkage map and other physical and genetic maps of potato and the closely related species tomato. Primary anchoring of the DM genome assembly was accomplished by the use of a diploid segregating population, which was genotyped with several types of molecular genetic markers to construct a new ~936 cM linkage map comprising 2469 marker loci. *In silico* anchoring approaches used genetic and physical maps from the diploid potato genotype RH89-039-16 (RH) and tomato. This combined approach has allowed 951 superscaffolds to be ordered into pseudomolecules corresponding to the 12 potato chromosomes. These pseudomolecules represent 674 Mb (~93%) of the 723 Mb genome assembly and 37,482 (~96%) of the 39,031 predicted genes. The superscaffold order and orientation within the pseudomolecules are closely collinear with independently constructed high density linkage maps. Comparisons between marker distribution and physical location reveal regions of greater and lesser recombination, as well as regions exhibiting significant segregation distortion. The work presented here has led to a greatly improved ordering of the potato reference genome superscaffolds into chromosomal "pseudomolecules".

KEYWORDS
Solanaceae
genome
anchoring
scaffold
orientation
sequence-tagged sites
pseudomolecules
potato
genetic map
physical map

Genome sequencing of crop plants has become increasingly routine, primarily due to the reduction in cost and increase in throughput brought about by continuing advances in sequencing technologies. First reports on the whole-genome sequences of plants, such as *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000) and rice (International Rice Genome Sequencing Project 2005), were mainly accomplished with the use of clone-based (*e.g.*, "BAC by BAC") strategies. In this approach, a library of bacterial artificial chromosome (BAC) clones is mapped onto chromosomes by the use of molecular markers, the aim being to generate a clone-based physical map with a "minimum tiling path." This assures good genome coverage while-minimizing the sequencing effort. More recently, plant genome sequencing has been based on whole-genome shotgun approaches involving conventional Sanger sequencing, next-generation sequence technologies, or a combination of both (Hamilton and Buell 2012). The whole-genome shotgun approach does not require a physical map, and there is no preassumption of the position of the resulting sequence assemblies. Several research groups have developed "scaffolding" algorithms to assemble these typically short sequence contigs into larger constructs (Miller *et al.* 2010). However, because of the genome size and complexity of most crop plants, scaffolds typically remain unoriented and without chromosomal coordinates, despite being well annotated for gene content. A reference genome sequence requires that the products of the assembly process (contigs and scaffolds) be globally ordered and oriented to generate chromosomal pseudomolecules (PMs). In the absence of a clone-based physical map or genetic map of the reference sequenced genotype, this task is a significant and challenging one. One widely adopted approach has been to link the sequence assembly to a genetic map using the presence of mapped sequence-tagged site (STS) genetic markers (Green and Green 1991) in the genome sequence. For example, a set of 409 molecular markers was used to order 69% of the assembled 487 Mb grapevine genome along the 19 grape linkage groups (The French-Italian Public Consortium for Grapevine Genome Characterization 2007). The link between the genome sequence and its genetic maps is critical in moving between trait loci and candidate genes underlying such loci. Successful genetic anchoring of a plant genome sequence assembly with the use of maps developed in the reference-sequenced genotype depends on marker density and distribution, as well as map accuracy and resolution. Other approaches can also be implemented to augment the anchoring process, including comparative analysis with physical and genetic maps of closely related species.

The Potato Genome Sequencing Consortium (Potato Genome Sequencing Consortium 2011) has published the genome of the doubled monoploid *Solanum tuberosum* group Phureja DM1-3 516 R44 (hereafter referred to as DM). At the time the genome sequencing was initiated, DM did not have a physical map, nor was there any pre-existing genetic map for this genotype. Therefore, a genome-anchoring strategy was developed that included the generation of a segregating bi-parental mapping population involving DM as a parent, and generation of a dense STS-based genetic map. Other genetic mapping resources, such as the ultra-high density (UHD) map of diploid potato genotype RH89-039-16 (RH) (van Os *et al.* 2006), and the tomato-EXPEN 2000 genetic reference map (Fulton *et al.* 2002) were also used.

We describe for the first time in detail the generation of an integrated *de novo* genetic/physical map of potato and significant refinements to the previously published assembly. Our combined map orders the genome sequence into 12 chromosomal PMs corresponding to each of the 12 potato chromosomes and is linked to previously existing potato and Solanaceae mapping resources. The work represents the assimilation of various data types that required complex interpretation for correct ordering and orientation of super-scaffolds. This process involved considerable manual curation, driven largely by a novel approach for visualization of mate-pair sequences from large genomic clones (BAC and fosmid) and long insert 454 reads (20 kb and 8 kb). This allowed us to assign robust orientations to many superscaffolds and also enabled the inclusion of many super-scaffolds that remained unanchored when the reference genome sequence was published (Potato Genome Sequencing Consortium 2011). This resource will facilitate exploitation of the potato genome sequence for genetic analysis and crop improvement, and our approach can serve as a guide for others wishing to engage in genome sequencing of genotypes which lack physical or genetic maps.

MATERIALS AND METHODS

Genetic cross/population construction

A segregating diploid potato population (BC_1) derived from the reference sequence clone DM 1-3 516 R44 (DM) was developed. The homozygous DM clone ($2n = 2x = 24$) was generated by chromosome doubling of a monoploid ($2n = 1x = 12$) derived from a heterozygous accession of *S. tuberosum* Group Phureja (Paz and Veilleux 1999). A heterozygous diploid clonal accession (CIP 703825, referred to as D) belonging to the *Solanum tuberosum* diploid Andigenum Group Goniocalyx cultivar group (Spooner *et al.* 2007; Ovchinnikova *et al.* 2011) was crossed to DM. The direction of the cross (DM × D) was chosen because DM is male sterile. One of the resulting F_1 hybrids (DM/D, CIP 305156.17) was used as the stylar parent in a backcross with D as pollen parent. The mapping population comprising 180 backcross progeny clones (hereafter referred to as DMDD) was raised in the greenhouse for DNA extraction and pathogen testing and is also maintained pathogen-free *in vitro* (https://research.cip.cgiar.org/confluence/display/dm/Home) at the International Potato Center, Peru.

Plant material and genomic DNA extraction

Genomic DNA from 180 progeny clones of the mapping population and the pedigree parents was isolated by the use of standard protocols (Herrera and Ghislain 2000). DNA concentration was estimated with a TBS-380 Fluorometer (Turner BioSystems) with PicoGreen reagent using salmon sperm DNA at 500 ng/mL as a reference. All DNA samples were normalized to a final concentration of 250 ng/μL and distributed among members of the Potato Genome Sequence Consortium (PGSC) mapping group to perform multilocation genotyping by using diversity arrays technology (DArT), simple sequence repeat

doi: 10.1534/g3.113.007153
Manuscript received July 4, 2013; accepted for publication September 10, 2013

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.113.007153/-/DC1.
[1]These authors contributed equally to this work.
[2]Present address: The EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, United Kingdom.
[3]Present address: Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Rio Grande do Sul, Brazil.
[4]Present Address: Department of Molecular Biology and Genetics, Aarhus University, Slagelse DK-4200, Denmark.
[5]Corresponding author: Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, United Kingdom.
E-mail: glenn.bryan@hutton.ac.uk

(SSR), single-nucleotide polymorphism (SNP), and amplified fragment-length polymorphism (AFLP) markers.

Marker identification, development, and analysis

SSR markers: SSR markers were designed from an early draft of the assembled potato genome superscaffolds (DM assembly version 1). Markers were selected from a masked copy of the genome to avoid placement in repetitive DNA. In addition to these SSR markers (labeled PM), previously reported sets of SSRs from Stwax (potato *waxy* gene; Veilleux *et al.* 1995), STM (Milbourne *et al.* 1998), STI (Feingold *et al.* 2005), st_ (Tang *et al.* 2008a), and STG (Ghislain *et al.* 2009) were also used in linkage mapping. In total, 356 SSRs (Supporting Information, Table S1A) were tested for polymorphism. In brief, 5−25 ng of template DNA was added to polymerase chain reaction (PCR) mix containing 1.5−2.5 mM $MgCl_2$, 0.16−0.25 mM dNTP, 0.25−1.0 U Taq polymerase, with the following primer combinations; for acrylamide gel analysis, 0.2−0.25 μM forward primer, 0.2−0.25 μM reverse primer, plus 0.2 mM cresol red and 6% sucrose; for ABI3130lx Genetic Analyzer (Applied Biosystems), 0.2−0.25 μM reverse primer, 0.15−0.25 μM forward primer, 0.05−0.25 μM labeled (FAM (5-FAM (6-FAM) 5(6)-carboxyfluorescein), HEX (6-carboxy-1,4-dichloro-2′,4′, 5′, 7′-tetrachlorofluorescein), NED, or PET) forward primer; for 4300 LI-COR DNA Analyzer (LI-COR Biosciences), 0.2 μM or 22 pM forward primer, 0.2 μM or 15 pM reverse primers, 25 pM 700 or 800 IRDye labeled M13 forward primer. PCRs were conducted under optimized conditions: in brief, 4 min denature at 94°, 35 cycles of 30 sec at 94°, 30 sec at T_a (annealing temperature determined experimentally for each SSR primer combination), 30 sec at 72°, 1 cycle of 4 min at 72°; or 3 min denature at 94°, 36 cycles of 15 sec at 94°, 30 sec at 58−52° with touchdown of −0.5° for first 12 cycles, 30 sec at 72°, 1 cycle of 5 min at 72°; or 4 min denature at 94°, 30−33 cycles of 1 min at 94°, 1 min at T_a, 1 min at 72°, 1 cycle of 4 min at 72°. SSRs were resolved either by denaturing acrylamide gel electrophoresis and silver staining according to Creste *et al.* (2001), capillary electrophoresis following standard procedures for the ABI3130lx Genetic Analyzer using Genscan 400 ROX (6-carboxy-X-rhodamine) or Genscan 500 LIZ size ladder, or by electrophoresis on the 4300 LI-COR DNA Analyzer system (LI-COR Biosciences) using the LI-COR IRDye 50–350 bp size standard. Polymorphic markers were scored directly from silver stained gels; using GeneMarker 1.4 (SoftGenetics, State College, PA; www.softgenetics.com), GeneMapper 4.0 (Applied Biosystems) or Genographer (www.genographer.com) for ABI3130 lx; or the SAGA Generation 2 software (LI-COR, USA), and Cross Checker v.2.9.1 (Buntjer 1999) for LI-COR. SSRs were scored, where possible, as codominant markers, and if this was not possible, as dominant markers.

SNP markers: A custom filtering pipeline was developed to select 1920 SNPs from a set of 69,011 high-confidence SolCAP SNPs (Hamilton *et al.* 2011) that were incorporated into five 384-plex (5 × 384) Illumina GoldenGate oligonucleotide pool assays (OPAs; Fan *et al.* 2003), hereafter referred to as POPA (potato OPAs). Hamilton *et al.* (2011) identified these SNPs by comparing RNA-Seq and EST sequences from six potato cultivars (Atlantic, Premier, Snowden, Bintje, Kennebec, and Shepody) to the draft DM potato reference genome. Our filtering pipeline involved finding nonrepetitive positions on the DM assembly, avoiding overlapping SNPs that may have interfered with the Illumina SNP genotyping assay, and striving to cover the genome as fully as possible. In addition, a POPA containing SNPs derived from pre-existing potato ESTs in the public databases was also designed and used. Table S1B shows details of 2304 SNPs, derived from pre-existing potato ESTs (POPA1) and SolCAP markers (POPA2-6) used in the study. Genotyping was performed using an Illumina BeadXpress platform following the recommendations of the manufacturer (GoldenGate Genotyping Assay, Illumina VeraCode Manual, VC-901-1001). All reagents, unless stated otherwise in the standard protocol, were provided by Illumina. The data files were processed and genotypes called using Genome Studio software.

AFLP markers: AFLP analysis was performed according to the procedures described by Vos *et al.* (1995) using the restriction enzyme combination *Eco*RI and *Mse*I. AFLP fragments were separated on a LI-COR 4300 DNA Sequencer (LI-COR Biosciences) using 4.5% polyacrylamide denaturing gels (acrylamide:bisacrylamide, 19:1) as described in the user manual. The LI-COR size standard ladder was loaded into each lane to facilitate the semiautomatic analysis of the gel and the sizing of the fragments. The names of the markers indicate the enzymes used, the selective nucleotides, and the size of the fragment; for instance, EACTMAAC_205.0 is an AFLP marker derived from a primer combination with the enzymes *Eco*RI and *Mse*I, selective nucleotides ACT and AAC, and a mobility that corresponds to a fragment with an estimated size of 205 bp. Polymorphic bands were manually scored following the intensity degree and the parent backcross pattern. The details of the enzyme combinations, selective nucleotides, and adapter sequences are provided in Table S1C.

DArT markers: Representations from 180 DMDD progeny clones and the pedigree parents (DM, DM/D, D) were obtained by subjecting DNA from each clone to double restriction enzyme digestion (*Pst*I/*Taq*I) and ligation to *Pst*I adaptors for reducing genome complexity followed by PCR amplification for preparation of targets (Wenzl *et al.* 2004). Cy3-labeled representations (targets), mixed in an ExpressHyb buffer containing cy5-labeled polylinker fragment of the plasmid used for library preparation (as a reference), were denatured and hybridized to a high-resolution potato genotyping array containing 7680 DArT probes (Sliwka *et al.* 2012). After overnight hybridization at 62°, arrays were washed and scanned with 20 μm resolution at 543 nm (cy3) and 488 nM (FAM) on a LS300 confocal laser scanner (Tecan, Grödig, Austria) to detect fluorescent signals emitted from the hybridized fragments. The data from the scanned images were extracted and analyzed using the DArTsoft 7.4 software (Diversity Arrays Technology P/L, Canberra, Australia). The logarithm of the ratio between the two background-subtracted averages of feature pixels in the cy3 and cy5 channels (log2[cy3/cy5]) was used as a measure of the difference in abundance of the corresponding DNA fragment in the two representations hybridized to the array. The log2[cy3/FAM] and log2[cy5/FAM] values, which are approximate measures of the amount of hybridization signal per amount of DNA spotted on the array, were used for quality-control purposes. The unique signal pattern obtained by hybridizing each sample pair (individual clone and reference) to the genotyping array was recorded as "0" or "1." All DArTs were sequenced and are available from Spud DB site (http://potato.plantbiology.msu.edu/); the detailed methodology is published on the Diversity Arrays Technology website (http://www.diversityarrays.com).

Linkage map construction

The SSR, SNP, AFLP, and DArT genotyping data for 180 DMDD progeny clones were combined and screened for polymorphic markers. JoinMap4 (Van Ooijen 2006) was used both to assign markers to linkage groups and to order markers within linkage

groups. The backcross parents and offspring were coded according to the cross-pollinated (CP) population type (outbreeder full-sib family after two independent meioses). A female-male combined DMDD map was generated that included markers informative in one or both parents. Linkage groups were formed using the Independence LOD parameter under "population grouping" with a range from 2 to 15. Before grouping and ordering markers within linkage groups, loci or progeny clones with ≥20% missing values were removed along with all identically segregating loci. The regression mapping algorithm with modified settings (recombination frequency threshold < 0.49, LOD threshold > 0.01) was used to order loci within each linkage group. All linkage groups were subjected to three rounds of mapping. Recombination frequencies were converted into map distances using the "Kosambi" mapping function.

Locating STS markers on the DM assembly

STS markers were aligned to the reference genome assembly using SSAHA2 (Ning *et al.* 2001) or BLAST. The total set of alignments was processed as follows. First, alignments caused by short repetitive sequences were removed using a custom depth/coverage filter. In detail, any alignment covering a region of the query or target sequence that overlapped with five or more other competing alignments in that region was removed if this depth threshold was exceeded greater than 20% or more of the alignment length. In this way alignments spanning short repeats were not penalized, but alignments largely composed of likely repeats were removed. Second, short alignments were grouped by sequence into "hits" that allowed for indels. Third, where applicable, the relative distance and orientation of the forward and reverse reads for the marker was taken into consideration. Pairs of forward and reverse reads with an incorrect orientation or implausible separation were removed. Finally, only markers with a unique, high-scoring alignment position on the genome assembly were selected as anchor points in the physical map. The final positions of all the STS markers (SSRs, SNPs, and DArTs) are provided in Table S2.

Integration of additional sequence-based and physical resources

DM BAC- and Fosmid-end sequences, RH BAC-end sequences, and tomato BAC- and Fosmid-end sequences were aligned to the DM superscaffolds using SSAHA2 (Ning *et al.* 2001). The resulting alignments were filtered as described previously. Roche 454 Paired-end (PE) reads from 14- and 20-kb insert-size libraries from DM, representing 0.7 and 1.0 Gb of raw data, respectively, were aligned to the superscaffold sequences using Newbler (Margulies *et al.* 2006) with all the default settings. Unsequenced BAC clones from the RH physical map (de Boer *et al.* 2012) were positioned on the superscaffolds using BLAST alignment of their whole-genome profiling (WGP) sequence tags. For each BAC, the alignment hits of the individual 25 nt tags were processed to retain only unique hits. The aligned BAC clones that carried AFLP markers provided the link between the DM superscaffolds and the RH UHD genetic map (van Os *et al.* 2006). In addition, sequenced RH BAC clones and RH BAC-end sequences were used for anchoring and scaffolding of the DM sequences. Finally, sequences from the available tomato PMs (v2.40, The Tomato Genome Sequencing Consortium 2012) were aligned using ATAC (Istrail *et al.* 2004).

Manual scaffolding using the "link-peak" strategy

All paired-end and mate-pair (PEMP) sequence data that could be reliably mapped to the DM superscaffolds were combined to compute a composite directional link-score across each superscaffold. In detail, the link-score combined PEMPs that had unique, high-scoring alignments for both ends of each mate pair sequence, but with the two end sequences aligning to different non-adjoining superscaffolds. A reciprocally high link-score between the ends of a pair of superscaffolds indicated a probable scaffolding link between them. The composite directional link-score is calculated in a sliding window along the length of a superscaffold (the source) as follows:

1. All mate pairs with one end aligning in that window and the other corresponding mate pair end reliably mapping to another superscaffold (the target) are selected. These are designated as unsatisfied mate pairs.
2. These mate pairs are grouped according to the target superscaffold.
3. For each target superscaffold group, a score is calculated by summing the value for each mate pair in that group (see below for details of how the value is determined).
4. The link-peak score is the greatest score of all the target groups.

Different link-score values were empirically assigned to the different PEMP sequence libraries, with greater scores assigned to DM based libraries over RH and tomato-based libraries and greater values given to longer sequences that have more accurate alignments. In addition to accumulating link-evidence from consistent unsatisfied PEMPs, a noise-score was calculated for unsatisfied PEMPs that suggested links to multiple different target superscaffolds. The noise score allowed spurious, high-scoring links caused by repeats to be identified. In this way the evidence for links between pairs of superscaffolds could be conveniently described as a continuous value in wiggle format (https://www.genome.ucsc.edu/goldenPath/help/wiggle.html), which allows for visualization as tracks in GBrowse, alongside genetic and physical evidence from other sources.

Visualization of integrated genetic and physical map

The integrated genetic and physical maps of the DM genome were visualized with the software 'DMAP' (D. M. A. Martin, unpublished data). The figures produced by the DMAP software take as input the accessioned golden path (AGP) file describing the PM architecture, a GFF file describing the sequence positions of the markers on the superscaffolds, and the JoinMap output file from linkage mapping for each linkage group. As there are many more markers than those that can be coherently visualized on a printed figure, DMAP employs a selection and layout algorithm where only a user determined maximum number of labels are displayed.

DM chromosome idiogram figures were reproduced from the potato reference genome publication (Potato Genome Sequencing Consortium 2011) and were aligned by orienting the short arms toward the start of the PM sequence, except for chromosomes 5 and 11, where the PM sequence begins in the long arm (Tang *et al.* 2009; Potato Genome Sequencing Consortium 2011).

Identification of centromere positions and pericentromeric regions

Centromere positions were determined with the sequence information provided by Gong *et al.* (2012). For chromosomes 4, 6, 9, 10, 11, and 12, the DM superscaffolds covering the centromere locations were identified from the major peaks in the CENH3 chromatin immunoprecipitation sequence read plots on the DM V2.1.10 PM sequences. Satellite repeat analysis was performed by searching for the repeats in

the DM sequence at http://yh.genomics.org.cn/potato/search.jsp and by evaluating the repeat coverage through dot plot alignment of candidate DM sequences with the repeat sequence. In addition, centromere positions were also indirectly inferred from the marker density in RH UHD genetic map (van Os *et al.* 2006).

The revised physical positions of all of the Illumina Potato 8303 Infinium array SNPs, reported by Felcher *et al.* (2012) using their customized version (2.1.11) of potato reference PMs, were obtained for the latest version (4.03) of PMs (Table S3). Graphs depicting the progression of genetic distance and recombination rate *vs.* physical distance were calculated for all of the SNPs included in the current PMs and D84 and DRH genetic maps, using the MareyMap package (Rezvoy *et al.* 2007). The pericentromeric heterochromatin regions of the DM PMs were identified in these plots from the absence of genetic recombination between the SNP markers in such regions. In addition, AFLP markers from the marker-dense pericentromeric bins of the RH genetic maps were used to define heterochromatin boundaries in the PMs (Park *et al.* 2007), especially in cases where the genetic maps of Felcher *et al.* (2012) offered limited resolution.

BAC assembly and comparison with PMs

A total of 96 DM BACs spanning scaffolding gaps on chromosome 4 were selected (using DM BAC-end hits; Potato Genome Sequencing Consortium 2011). The BACs were picked from the library and end-sequenced to verify correct selection. Eighty-two verified BACs were further processed and grouped into six normalized pools as well as a composite master pool containing all 82 BACs. Each of the six BAC pools was subjected to Roche 454 single-end shotgun sequencing and the master pool to 3-kb PE sequencing. Single-end data for each pool were combined with the PE data and were assembled together using the Newbler GSAssembler (Margulies *et al.* 2006). The sequences were deposited in the EBI Short Read Archive (accession number: ERP000934).

Candidate BAC scaffolds containing BAC-end sequences were identified with BLAST, filtering hits with a minimum match length of 400 bases and bit score exceeding 700 before manual curation. BAC scaffolds were matched to DM genomic superscaffolds with MUMmer (Kurtz *et al.* 2004). Matching regions were filtered to retain only matches longer than 1000bp with >97% identity. Data were expressed graphically with matches as edges and BAC end sequences, superscaffolds and BAC scaffolds as nodes using the graphical exchange format. Code was written in Python with the pygexf library and visualization performed with Gephi (http://www.gephi.org). In addition, BAC ends were linked by a BAC label as a node. Assemblies which linked superscaffolds with sequence data could then be readily observed as cycles containing a BAC label in the graph. BAC-oriented GFF files were generated and visualized with R.

RESULTS AND DISCUSSION

DM genome assembly: a brief summary

The potato nuclear genome involved generation of ~96.6 Gb of raw sequence, which assembled into 66,254 "superscaffolds" comprising a net sequence assembly of 727 Mb, 117 Mb less than the estimated genome size of 844 Mb. Superscaffold length is inversely proportional to the numerical value in the name of each DM superscaffold (DMB), where the largest DMB (7.1 Mb) bears the ID "PGSC0003DMB000000001" and the smallest (100 bp) "PGSC0003DMB000066254." Approximately 94% of the assembled genome is nongapped sequence and more than 90% of the genome (N_{90}) is represented by 622 superscaffolds that are equal to or larger than 0.25 Mb. The anchoring strategy preferentially targeted the larger superscaffolds. At the time of publication 649 superscaffolds equaling 623 Mb (86%) of the assembled genome and 90% of the 39,031 estimated genes were anchored (Potato Genome Sequencing Consortium 2011). Draft PMs for the 12 chromosomes had been constructed but superscaffolds were mostly un-oriented. Since the original publication, continuous efforts have been made to perform further anchoring and orientation of the DM superscaffolds in order to generate the revised and improved genome PMs presented here (version 4.03).

Genetic analysis of the mapping population

The DMDD mapping population was genotyped for AFLP, SSR, SNP, and DArT markers. Twenty two AFLP primer pairs (*Eco*RI/*Mse*I) amplified 213 detectable fragments. A total of 356 SSR loci were assayed. Of 2304 POPA SNPs and 7680 DArTs interrogated, 2160 and 2174 yielded genotype data, respectively. The compiled set of 4903 markers was screened for presence of polymorphism, data integrity, and concordance between parental and progeny genotypes, as well as meeting the missing data threshold (<20%) and other standard quality control checks. These data filtering and quality measures resulted in considerable reduction in the total number of markers used for linkage mapping to 2597, which comprised 187 AFLPs, 234 SSRs, 367 SNPs, and 1809 DArTs. After excluding co-segregating markers, we used a subset of 1864 uniquely segregating loci for linkage grouping; 1751 unique loci were incorporated into a combined parental linkage map with the 12 expected linkage groups, whereas the remaining 113 remained unmapped. The 12 chromosomal linkage groups span 936.2 cM with an average marker spacing of 0.54 cM per interval. The individual linkage groups ranged in size from 62.9 cM (Chr11) to 101.8 cM (Chr03). A combination of the use of previously mapped SSR markers (Veilleux *et al.* 1995; Milbourne *et al.* 1998; Feingold *et al.* 2005; Tang *et al.* 2008a; Ghislain *et al.* 2009) and other available resources such as the RH genetic map (Van Os *et al.* 2006), the RH WGP map (de Boer *et al.* 2012) and the tomato-EXPEN 2000 map (Fulton *et al.* 2002) allowed orientation and assignment of all 12 linkage groups to their respective chromosomes. Table 1 shows the summary statistics of linkage mapping in the DMDD cross.

Departure from Mendelian segregation has been observed frequently in potato crosses. Markers showing segregation distortion were not excluded from the mapping process and most could be mapped to their appropriate linkage groups. The frequency of segregation distortion was highly variable among different chromosomes with the most significant distorted regions observed on chromosomes 1 and 4. Previous potato mapping studies have also shown varying levels of segregation distortion (Gebhardt *et al.* 1991, Felcher *et al.* 2012). Figure S1 shows genome-wide distribution of levels of segregation distortion for all STS markers used in DMDD.

Linkage map–based (direct) anchoring

The linkage map of DMDD is predominantly composed of STS markers. The primary map-based anchoring strategy involved locating these sequence-based markers in the DM superscaffolds. SNPs and previously unpublished SSR markers (prefixed with "PM") used in the DMDD linkage map were designed *a priori* against genome superscaffolds so their unique positions in the relevant superscaffolds were known. The positions of DArT and previously reported SSRs were determined using the bioinformatics alignment and filtering pipeline illustrated in Figure 1.

Co-segregating markers removed during linkage map construction were included in the anchoring process as such genetically redundant markers represent distinct, but physically linked sites in the genome.

Table 1 Distribution of 1751 markers comprising four different classes across the 12 chromosomes in the DMDD population, with the concomitant map and interval lengths (cM) for each chromosome

Chr[a]	Mapped Markers[b]	Map Length, cM	Interval Spacing, cM/interval[c]
01	201	93.0	0.46
02	221	77.4	0.35
03	134	101.8	0.77
04	143	99.7	0.70
05	107	64.1	0.61
06	134	70.5	0.53
07	108	67.1	0.63
08	176	67.8	0.39
09	152	87.9	0.58
10	144	68.9	0.48
11	108	62.9	0.59
12	123	75.2	0.62
All	1751	936.2	0.54

SSR, simple sequence repeat.
[a] Based on the SSRs mapped in previous studies and further confirmed by using *in silico* approaches.
[b] Excluding 718 co-segregating markers; when the segregation pattern of two or more markers was identical, only a single marker per set of identical markers was retained to generate the maps; 128 ungrouped markers (including 15 unassigned co-segregating markers) that did not fit any linkage group were also excluded.
[c] Calculated as the map length divided by the number of intervals (mapped markers minus 1, for "total" it is mapped markers minus 12).

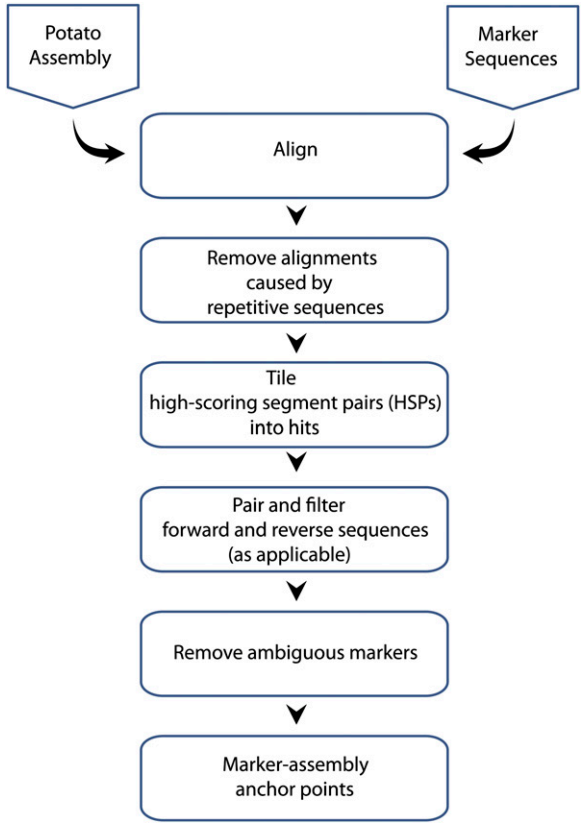

Figure 1 Pipeline for anchoring of markers to the potato genome assembly.

The complete set of STS markers was filtered for unique and unambiguous marker-assembly sequence alignments as described. The combined sequence and genetic map coordinates for these unique STS markers were used to assign and order superscaffolds for constructing a framework physical map. The integrated genetic and physical anchoring strategy is shown in Figure 2. Using this strategy, we anchored 1730 (1305 DArTs, 345 SNPs, and 80 SSRs) of the 2292 mapped, including co-segregating, STS markers to their unique positions on the DM superscaffolds. This approach anchored 54.2% (394 Mb) of the DM genome assembly arranged into 334 superscaffolds (Table 2). The proportion of genetic markers anchored on the genome sequence from each marker-category was 96% (SNPs), 28% (SSRs), and 76% (DArTs). Mapped AFLP fragments were not used in the anchoring process, due to a lack of sequence information. Table S2 contains genomic positions for all the STS markers used in the study. Genetic and physical coordinates for the DMDD mapped markers, including 718 co-segregating markers, are provided in Table S4.

In silico approach–based (indirect) anchoring

The DMDD-based framework physical map was extended by integrating two additional sources of syntenic map data, from potato and tomato, respectively. First, superscaffolds anchored using the RH UHD genetic and physical maps (van Os *et al.* 2006; de Boer *et al.* 2012) were added. Second, 2,604 sequence-based markers from the tomato-EXPEN 2000 derived maps, which are closely collinear with potato (Tanksley *et al.* 1992; Fulton *et al.* 2002; The Tomato Genome Sequencing Consortium 2012), were used to add superscaffolds. In the case of RH, sequence anchoring was derived from the AFLP- and WGP-based hybrid RH physical map (de Boer *et al.* 2012) as well as by direct alignment of RH BAC end sequences and fully sequenced RH seed BACs to the DM sequence. In both cases, the (proxy) marker sequences were aligned to the DM assembly using BLAST, adopting stringent matching criteria. The results were processed into reliable genetic anchor points as described previously for the DM markers.

The RH- and tomato-based *in silico* anchoring strategies independently anchored 470 (527 Mb, 72.5%) and 402 (417 Mb, 57.4%) superscaffolds, respectively (Table 2). Figure 3 shows the superscaffold anchoring summary for both the linkage (DM map) and the two *in silico* (RH and tomato maps) approaches. The total set of 649 superscaffolds anchored in at least one map was integrated hierarchically, starting with the DMDD-based framework map, placing additional superscaffolds using first the RH and then tomato assignment. The hierarchical 'alignment' of the maps is described below.

Construction of chromosome-scale PMs

Following anchoring, the superscaffolds were ordered into chromosome-scale PMs in a hierarchical process using genetic, sequence and physical map data. The process is broken into two stages.

Stage I: In the first stage the STS markers from the DMDD genetic map were aligned to the DM superscaffolds and used to construct the "backbone" PMs. Additional sequence-linked and sequence-based markers from the RH and tomato genetic maps were subsequently used to add superscaffolds into the DM backbone PMs (Figure 2). Superscaffolds that were anchored in multiple maps were used as reference points to align the genetic positions in the three different maps. Superscaffolds were added into 'gaps' in the backbone PMs where the positions indicated by the RH and tomato markers were in agreement with the positions initially established by the DMDD map data. The known set of chromosomal inversions on chromosomes 5, 6, 9, 10, 11, and 12 between potato and tomato (Tanksley

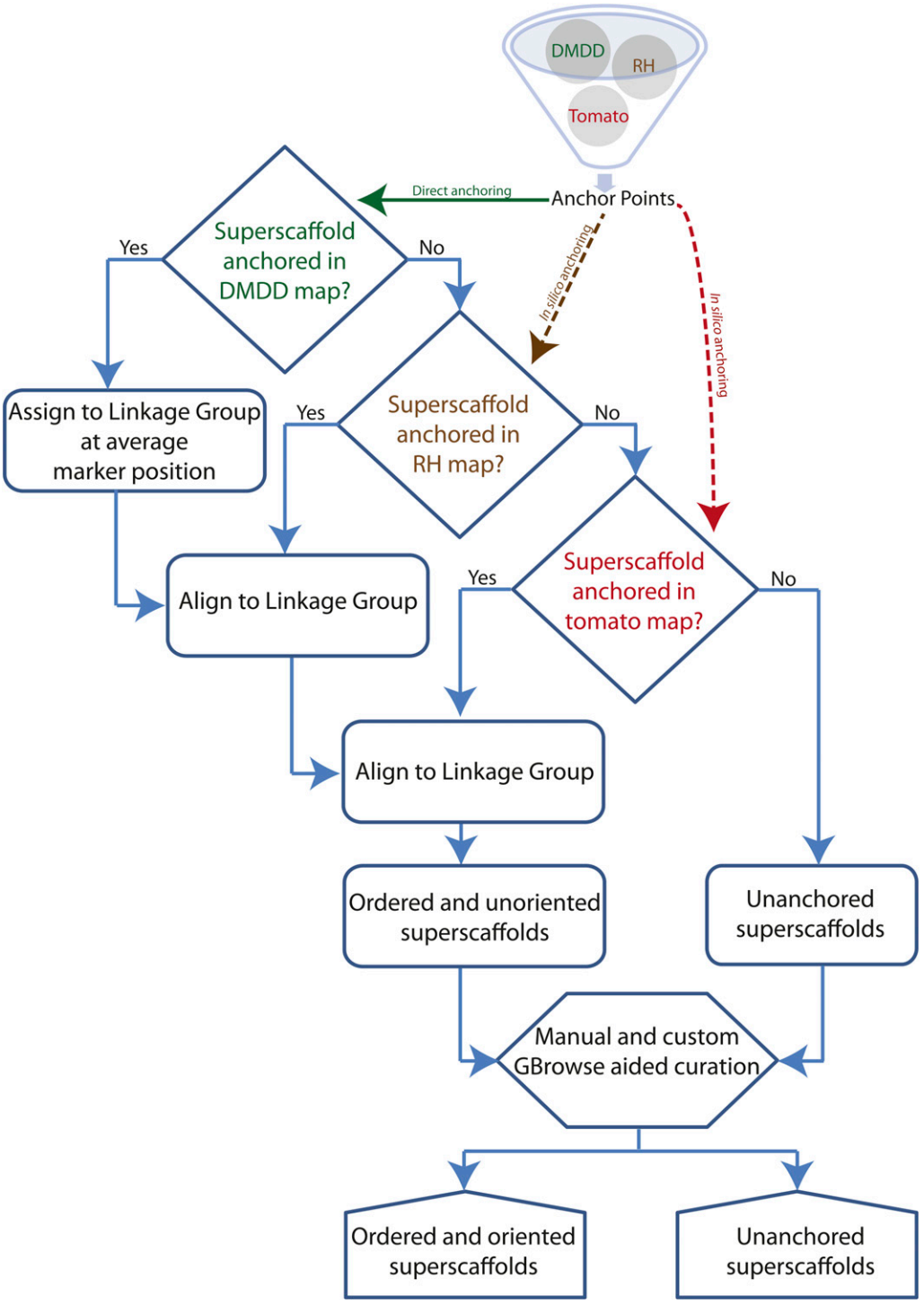

Figure 2 Step-wise linkage group assignment and ordering of DM superscaffolds using genetic-anchoring information successively from the DM, RH, and tomato genetic maps.

et al. 1992; Iovene *et al.* 2008; Tang *et al.* 2008b) were taken into account when aligning the different genetic maps.

Generally the different anchoring approaches provided direct support for each other with respect to the relative placement of superscaffolds in the PM. With an optimal alignment/agreement for the superscaffold order among the three different maps used for anchoring, 294 of 374 superscaffolds present in at least one map were found to be in the same order as in the other two maps. In some instances, we observed that ordering of superscaffolds derived using RH and tomato maps was inconsistent with that obtained from the DMDD genetic map. The observed differences could be due to many factors, including technical issues such as mapping or assembly errors or biological properties, such as previously unknown structural differences between the compared genomes. However, given the size and complexity of the potato genome, it is encouraging that the placement of 79% of the superscaffolds was corroborated by the different methods employed.

Although superscaffolds were integrated into genomic blocks at this stage, they were unoriented and, due to the difficulty of aligning genetic maps, largely unordered at the chromosome level. To add, orient and refine the order of superscaffolds into an AGP for

Table 2 Anchoring statistics by chromosome for the three different physical maps, *de novo* (DM) and *in silico* (RH and tomato)

Chromosome	DM Map			RH Map			Tomato Map		
	DMB Anchored	Cumulative Length, Mb	No. of Markers[a]	DMB Anchored	Cumulative Length, Mb	No. of Markers	DMB Anchored	Cumulative Length, Mb	No. of Markers
01	39	45	162	69	80	208	43	41	271
02	35	43	175	35	43	120	33	40	233
03	19	24	108	28	27	73	41	45	194
04	34	47	138	51	57	168	40	39	174
05	20	27	74	33	45	137	25	30	112
06	29	34	108	44	46	119	34	34	133
07	26	24	89	35	39	122	32	31	136
08	32	32	152	24	23	57	40	32	129
09	27	28	109	34	33	91	40	39	136
10	31	38	106	34	44	102	26	32	110
11	20	26	113	36	38	110	22	26	116
12	22	26	72	47	52	164	26	28	109
Total	334	394	1406	470	527	1471	402	417	1853

DM, doubled monoploid reference clone; RH, RH89-039-16; DMB, DM superscaffold.
[a] Only markers mapped in DMDD and uniquely and reliably anchored to DM assembly are included.

Figure 3 Summary of DM genome assembly anchoring using three different map resources. The number of uniquely and jointly anchored superscaffolds for each resource is given in the appropriate intersection. Cumulative size (Mb) of superscaffolds anchored in each category is shown in parenthesis. The total number of 649 anchored superscaffolds represents 623 Mb of the assembled DM potato genome. Figure updated from the Potato Genome Sequencing Consortium (2011).

constructing chromosome-scale PMs, a separate process was implemented, as described below.

Stage II: To orient the DM superscaffolds, and to further refine the DMDD linkage map-based PMs, sequence and physical data from a variety of sources were combined as described in the *Materials and Methods* section and visualized on a standard GBrowse installation (Figure 4). Custom sequence features were created representing high scoring intersuperscaffold links, allowing the user to "click-and-walk" along the physical evidence from superscaffold to superscaffold in GBrowse. To aid this visualization, the processed RH WGP and tomato alignments, including the aligned sequence markers from the genetic maps used in stage I, were added to GBrowse as additional sequence feature tracks.

Using this integrated visualization tool, we performed three important types of manual improvements to the stage I PMs: (1) scaffolding links were used to provide the relative orientation of superscaffolds, (2) adjacent superscaffolds not previously included in the integrated genetic/physical map were added, and (3) errors in the assembly were identified. These manual improvements were mainly carried out for the euchromatic (gene-rich) regions and for the euchromatin/heterochromatin borders. In addition to orientating the majority of the anchored superscaffolds, the "link-peak" walk strategy combined with manual curation led to the incorporation of an additional 277 previously unanchored superscaffolds into the PMs.

During this process 67 chimeric superscaffolds were identified. Of these, 62, 3, and 2 superscaffolds were revealed to have one, two, and

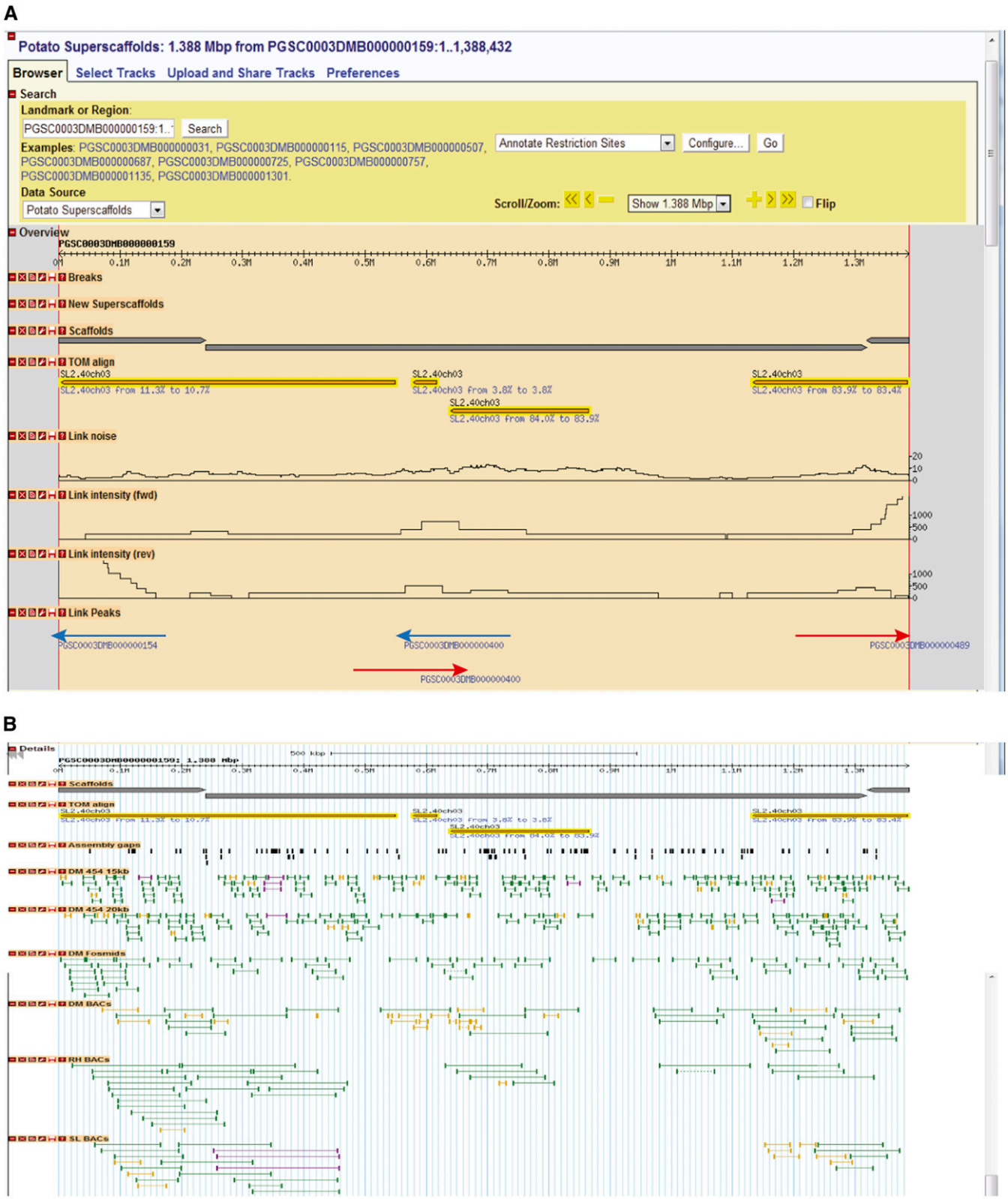

Figure 4 Depiction of "Link-peak" walk strategy taking superscaffold PGSC0003DMB000000159 as an example. (A) Custom GBrowse "Link-peak" intensity track features (shown as red and blue arrows) provided ordered navigation through superscaffolds using the aggregated PEMP. Link peaks to the right (red arrow) indicate "suggested path" downstream of the AGP, whereas those to the left (blue arrow) indicate converse. Reversal of this trend indicates a negative strand for the superscaffold in question. Traversing from one superscaffold to another by taking leads from these 'Link-peak' intensity tracks assisted in manually curating all 12 PMs. (B) Visualization of the underlying PEMP data.

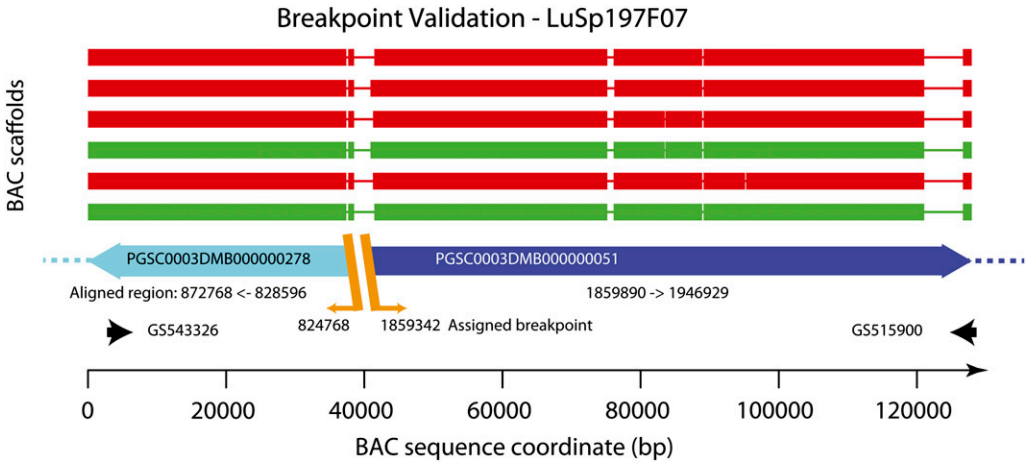

Figure 5 Assembled BAC sequence for LuSP197F07. Each scaffold assembly is derived from PE sequences of a combined pool of 82 DM BACs (spanning scaffolding gaps on chromosome 4) and single end sequence at greater read depth from one of the six subpools derived from the same BACs. The assemblies show a direct sequence running from PGSC0003DMB000000278 (− orientation, full length, cyan) through into PGSC0003DMB000000051 (+ orientation, blue) in accordance with the AGP and fully validating the decision to split PGSC00003DMB0000000278 at position 824768 and to split PGSC0003DMB000000051 at position 1859342 as indicated in the AGP file. Regions of good alignment (>98% identity, >1000 bases) are indicated as thick lines. Thin lines indicate no good alignment between the superscaffold and BAC sequences. The BAC end sequences are labeled with their Genbank IDs and are indicated at each end of the plot by black arrows. Breakpoints in the BAC sequences are indicated by orange diagonal lines and annotated with the assigned breakpoints coordinate from the AGP.

three misassembly locations, respectively, where false sequence joins had occurred. Many of these errors explained incongruities initially observed in the construction of the backbone PMs from the DMDD map (stage I). Chimeric superscaffolds were manually split and allocated to their respective positions in the PMs. For example, the sequence coordinates 1 to 1117982 bp of PGSC0003DMB000000002 were allocated to chromosome 4, whereas those from 1117983 to 6562806 bp were allocated to chromosome 5. These results further illustrate the utility of an integrated genetic and *in silico* anchoring based approach for refining and correcting genome assembly errors.

Included in the refinement process were dot plot alignments of DM chromosome PM sequences to pre-release and finished versions of the tomato genome sequence (The Tomato Genome Sequencing Consortium 2012). These alignments focused on the euchromatic regions and the adjacent heterochromatin border regions, where potato and tomato display homology in their sequences. The dot plot alignments to tomato made useful suggestions on how to place as yet unordered potato superscaffolds and superscaffold blocks, after which nearly always BAC end sequence links were identified in potato that confirmed the suggested orientation. Very occasionally, the potato PM description relied on the tomato alignment for placing potato sequence blocks in their presumed orientation, *e.g.*, from PGSC0003DMB000000729 to PGSC0003DMB000000835 at the top of chromosome 1 and from PGSC0003DMB000000692 to PGSC0003DMB000001163 in the south heterochromatin border on chromosome 8.

Inversions with tomato

The potato-tomato dot plot alignments explained the discrepancies that were found between the potato and tomato genetic maps. In the euchromatic regions and the adjacent heterochromatin border regions we collected the sequence positions of the 19 largest paracentric inversions (with a length of at least 0.3 Mb), which are listed in Table S5 and also indicated in the DM PM figures. Newly identified were, among others, a tandem inversion with minor additional rearrangements on potato chromosome arm 1L, a nested inversion on 2L, and an arm inversion on 8S. Furthermore, the known arm inversions on 9S and 11L were found to be tandem inversions, with the second inversion being located in the heterochromatin border. The chromosomal rearrangements on 2L have also been described by Peters *et al.* (2012), who presented a scenario involving four structural conversions between potato and tomato. However, our dot plot sequence alignment for this region is less complex and shows a single, smaller inversion inside a larger inversion. This nested inversion model requires only two structural conversion steps and remains compatible with the cytogenetic results of Peters *et al.* (2012).

No paracentric inversions were identified on chromosome 3. However, on the short arm, the tomato sequence differs from the potato sequence by a 7.0-Mb insertion, which is located at position 2.4 Mb in the DM chromosome 3 PM, and which runs from 1.3 to 8.3 Mb in the tomato SL2.40 assembly. In its center, this tomato insert has 4.2 Mb of sequence that is largely devoid of genes (http://potato.plantbiology.msu.edu/), while the start and end regions align with gene-containing potato sequence segments from region 42.0 to 50.4 Mb on the south arm of chromosome 3. Although these data suggest a translocation of sequences across the centromere, further investigation is needed to exclude sequence assembly errors.

Validation of link peak-based orientation strategy for chromosome 4

The strategy for PM construction and assembly correction was validated on chromosome 4 by targeted sequencing of 82 DM BAC clones that were selected to overlap candidate links as well as 10 of the 15 putative chimeric superscaffolds mapped to this chromosome. Thirty-one BAC clones could be assembled with contigs which spanned multiple superscaffolds and provided full coverage between the BAC end sequence matches to the superscaffolds, both validating the assembly and providing direct evidence for all 10 chimeric breakpoints. Seven of these sequenced BACs allow the inclusion of further superscaffolds that had not previously been assigned to a PM, and one provides evidence for a superscaffold that had been erroneously included.

In addition to the complete assemblies described previously, most other clones could be assembled to a series of contigs which did not span multiple superscaffolds and which have not been included in the BAC pool assembly summary (Table S6). Details of the BAC analysis are given in the *Materials and Methods* section and a representative example validating a potential break-point in Chromosome 4 is illustrated in Figure 5. A list of putative erroneous superscaffold assembly

locations (breakpoints), and the BACs which provide validation for them are given in Table S7.

Demarcating centromeres and pericentromeric boundaries in the PMs

The putative centromere locations for 7 of the 12 potato chromosomes were identified in the PM sequences based on data published by Gong *et al.* 2012 (Table S8). Six centromere locations were identified from chromatin immunoprecipitated sequences. Of the seven published centromeric satellite repeat sequences (Gong *et al.* 2012), only the St24 repeat specific for the chromosome 1 centromere identified DM sequences with a high repeat copy number characteristic of centromeric regions. With the other six centromeric repeat sequences, we could not find reliable centromeric targets in the DM assembly because these sequences only identified locations with very few repeat copies, which sometimes occurred on a chromosome other than that expected from their designated centromeres.

Pericentromeric boundaries were deduced by comparing the SNP-based D84 and DRH genetic maps of Felcher *et al.* (2012) to the current version of PMs. For all chromosomes the typical pattern of distinctly reduced recombination in pericentromeric regions and increased varying recombination rates in euchromatic regions was observed (Figure 6). These patterns were used as the primary information source to demarcate putative pericentromeric regions in the PMs, and the boundaries of these regions were well supported, and where needed refined, by the RH genetic maps (van Os *et al.* 2006). Figure 7 and Figure S2 depict the centromere and pericentromeric locations in the PMs. The pachytene chromosome idiograms in these figures are adapted from Potato Genome Sequencing Consortium (2011).

Current status of the reference PMs

The genome anchoring, ordering, and orienting process, as described previously, led to the joining of 951 genome superscaffolds, or nonchimeric segments thereof, into 144 larger, contiguous sequence blocks, and enabled construction of an AGP assembly for the reference DM potato genome. These chromosome-scale PMs, version 4.03, contain 93% (compared with 86%; Potato Genome Sequencing Consortium 2011) of the assembled genome comprising 674 Mb in 951 superscaffolds and include 37,482 (~96%) of the 39,031 predicted genes. A total of 938 superscaffolds (655 Mb or ~90% of the assembled genome sequence) are assigned absolute or relative orientation within the PMs, whereas the remaining 13 superscaffolds (19 Mb) are assigned with a random orientation. For 279 Mb of superscaffold sequence blocks from the heterochromatin, the exact chromosome position and absolute orientation could not be determined. These partially unordered regions are marked yellow in the PM figures (Figure 7 and Figure S2). No attempts were made to estimate gap sizes between the superscaffolds, and in the PM sequences all superscaffolds are separated from each other by a fixed gap sequence of 50,000 Ns. The N_{90} of the DM potato genome assembly is 0.25 Mb and contains 622 superscaffolds, of which 28 (equalling 17 Mb, ~2% of the assembled genome sequence) remain unanchored. The longest anchored superscaffold is 7.1 Mb (PGSC0003DMB000000001; chromosome 1) and the longest unanchored superscaffold (PGSC0003DMB000000064) is 2.2 Mb. The increase in average N_{50} from 1.5 Mb to 4.1 Mb in DM version 4.03 (Table 3) further supports the enhanced quality of the constructed PMs. The current version of the PMs/AGP is provided in Table S9 and includes the list of unanchored (chromosome 0) and chimeric superscaffolds.

For visualizing the differences and improvements in the constructed PMs, we compared dot plots of the current PMs (ver 4.03) to the earlier version 2.1.11 (Figure 8). Superscaffold misplacements were apparent as horizontal or vertical shifts in parts of the alignments in all pairwise comparisons. The overall structural integrity of the constructed PMs is visible from the expected gradual transition from gene rich to gene poor regions which in turn are well complemented by the normal high repeat region density patterns in the pericentromeric locations gradually declining toward the gene rich euchromatic regions (Figure 6). The PMs along with integrated DMDD and RH genetic maps were visualized using DMAP as described in the *Materials and Methods* section. Figure 7 shows a representative illustration for chromosome 1 (chromosomes 2−12 are shown in Figure S2). Good correspondence between DMDD and RH genetic maps and the PMs was observed.

Although the DMDD map-based strategy was critical in providing the basic anchoring to the DM genome, it had its limitations. Certain superscaffolds lacked sufficient polymorphic STS markers for genomic anchoring and were possibly affected by homozygosity, segregation distortion or other issues (Figure S1). This mainly occurred in pericentromeric/heterochromatin regions (marked by dashed lines, Figure 7 and Figure S2), which generally displayed a sparse coverage with DMDD markers, possibly due to the customized marker design strategy that precluded the design of markers in highly repetitive, relatively gene poor regions. For example, SNPs were designed against coding regions using RNA-Seq data (Hamilton *et al.* 2011) and, thus, were mainly localized to gene-rich regions, which occupy a different "genomic space" to the gene-poor, high-repeat content regions (Figure 6). The DM-based "PM series" SSRs were designed from repeat-masked genome sequence to avoid placement in repetitive DNA. The DArT methodology also uses genome complexity reduction and has been shown to target the low copy fraction of a plant genome through judicious selection of certain restriction enzymes (Jaccoud *et al.* 2001). Thus, the unavoidable bias toward nonrepetitive sequences in the STS markers employed in the DMDD map resulted in many unanchored superscaffolds. This issue was resolved by using additional resources that we refer to as the *in silico* anchoring approach. For example, the large block of "orphaned" superscaffolds, not directly connected to the DMDD map, stretching from DMB 394 to DMB 705 (with the exception of DMBs 193, 15, 59, 100, and 200) on chromosome 1 (see Figure 7) was anchored by the evidence derived from the WGP/AFLP-based RH map and the tomato-EXPEN 2000 map and further extended by the "link-peak walk" strategy, illustrating the importance of the multi-layered anchoring approach adopted here.

Potato genomic resources are provided as tracks/features in the GBrowse for the DM genome (hosted at Spud DB site "http://potato.plantbiology.msu.edu/"). One such resource, widely adopted by the potato community, is the Illumina Potato 8303 SNP Infinium array (Felcher *et al.* 2012) released after our map was constructed. This SNP array was used by Felcher *et al.* (2012) to construct two genetic maps, both involving DM as the female parent. Although the homozygosity of DM precluded segregation of DM loci in these populations, they showed good congruence for most linkage groups to the prerelease version (a modified ver 2.1.10 latterly referred to as ver 2.1.11) of the DM PMs. Version 4.03 of the PMs provides an improved correspondence with the genetic maps of Felcher *et al.* (2012) (Figure 6). An updated annotation of the Illumina Potato 8303 SNP Infinium array is provided in Table S3. The DMDD genetic map and associated data files are available at http://solgenomics.net/, and include hyperlinks to the MSU Genome Browser. All of the supplementary data, wherever applicable, are available to download as GFF format files from Spud

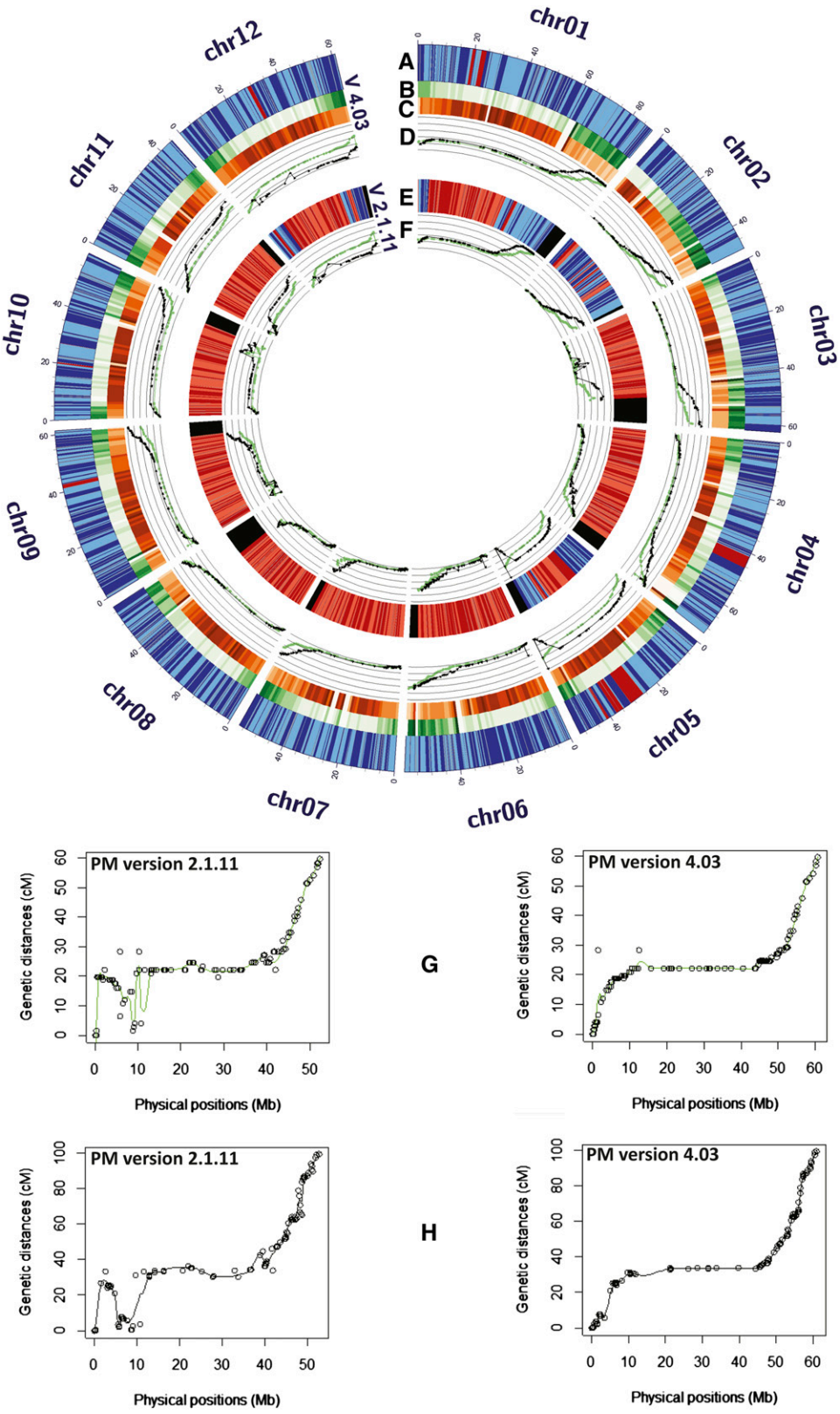

Figure 6 Enhanced accuracy of the current DM PMs. Panels A and E show anchoring of superscaffolds to the PM versions 4.03 and 2.1.11, respectively. Superscaffolds with known and unknown orientations are depicted in alternating shades of blue and red, respectively. Gaps in between the superscaffolds are marked in gray. Black areas in panel E represent unanchored superscaffolds (version 2.1.11) that were eventually anchored and ordered in PM version 4.03. Panels B and C show gene and repeat region densities, respectively, in 1 MB bins of PM version 4.03. Gene and repeat region densities ranges from 0 to >150 genes/MB and 0 to >900 repeats/MB, respectively. Panels D and F show the correspondence of the genetic maps (D84, green; DRH, black), adapted from Felcher *et al.* (2012), to PM versions 4.03 and 2.1.11, respectively. Graphs show the genetic (cM) positions plotted against the physical coordinates (Mb) for the SolCAP SNP markers; panels G (D84) and H (DRH) show elaborated examples of good correspondence from chromosome 9.

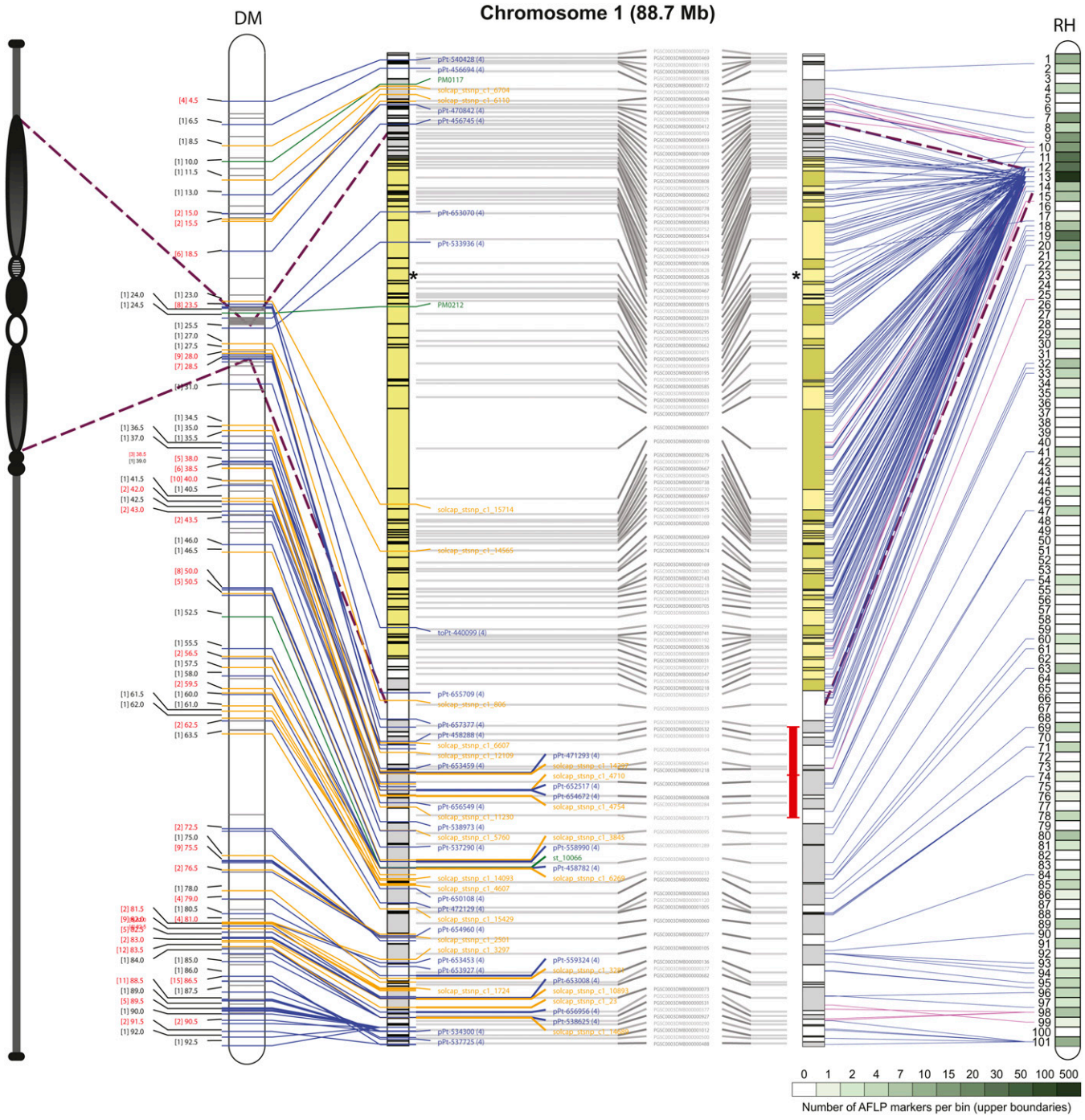

Figure 7 Illustration of the chromosome 1 PM integrated with the DM and RH genetic maps. STS and AFLP markers anchor sequence locations in the chromosome 1 PM to the DMDD and RH genetic maps, respectively. The AFLP marker positions in the PM were identified through sequence tag alignment of BAC clones from the RH WGP physical map. Superscaffolds comprising the PM are shown as alternating gray and white rectangular blocks. The layout of the PM for each of the genetic maps is shown separately but is identical with superscaffold IDs depicted in the middle. The pachytene idiogram is adapted from the potato reference genome publication (Potato Genome Sequencing Consortium 2011). The putative centromere region and pericentromeric/heterochromatic boundaries are demarcated by asterisk and dashed lines, respectively. Each DMDD marker type is color coded: blue = DArTs, yellow = SNPs, green = SSRs. Blue and magenta lines emerging from the RH genetic map represent AFLP anchors and the intensity of green color corresponds to the AFLP marker density per bin as reported by Van Os *et al.* (2006). Magenta lines represent AFLP markers with a relatively inaccurate mapping position on the RH genetic map, covering an interval of 5 or more bins. Regions in the central heterochromatin where superscaffold order and orientation are not completely resolved are indicated in yellow. Inversions with the tomato sequence are indicated with red interval bars.

Table 3 Improvements in DM PMs before and after execution of the link peak-based orientation strategy

Chr	Stage I[a] DMB Anchored		Stage II[b] DMB Anchored		Stage II[b] DMB Oriented[c]	
	No (Size in Mb)	N_{50}	No (Size in Mb)	N_{50}	No (Size in Mb)	Percentage
01	83 (79.7)	1.7	123 (82.6)	2.6	121 (79.8)	96.6
02	51 (45.0)	1.3	68 (45.3)	2.2	68 (45.3)	100.0
03	53 (45.3)	1.6	103 (57.2)	4.3	103 (57.2)	100.0
04	73 (60.9)	1.2	120 (66.3)	2.9	119 (62.1)	93.7
05	41 (44.8)	1.7	52 (49.5)	2.9	47 (40.4)	81.6
06	63 (54.0)	1.4	90 (55.1)	2.7	90 (55.1)	100.0
07	52 (50.6)	1.8	78 (52.9)	7.2	78 (52.9)	100.0
08	51 (41.6)	1.2	91 (52.4)	4.9	91 (52.4)	100.0
09	61 (50.6)	1.2	86 (57.3)	8.3	85 (55.9)	97.7
10	50 (51.4)	1.5	77 (56.0)	4.1	74 (55.4)	99.0
11	35 (34.4)	1.4	60 (42.5)	5.7	60 (42.5)	100.0
12	61 (58.5)	1.5	77 (57.4)	1.9	76 (56.0)	97.7
Total	674 (616.8)[d]	1.5[e]	1025[d,f] (674.4)[d]	4.1[e]	1012[d,f] (655.1)[d]	97.2[e]

DM, doubled monoploid reference clone; PMs, pseudomolecules; DMB, DM superscaffold.
[a] Refers to the status of PMs before execution of the "Link-peak" walk strategy.
[b] Refers to the status of PMs after execution of the "Link-peak" walk strategy.
[c] Only attempted at stage II.
[d] Total.
[e] Average.
[f] Chimeric superscaffolds have been included more than once (net number of DMBs anchored = 951).

DB site "http://potato.plantbiology.msu.edu/". The potato GBrowse including all of the hosted genomic resources/tracks/features have also been updated to the latest version (PM 4.03) of the DM PMs.

Conclusions

The integrated genetic and physical reference map presented here comprising nearly 2500 markers, which are mostly STS, provides a platform for exploiting the potato reference genome. The most obvious and immediate application is the ability to position any sequence-based marker locus to a precise location in the DM genome. This will revolutionize trait analysis, although progress will be dependent on the complexity of the trait concerned, population size, replication and accuracy of phenotypic data and other factors that impinge on map resolution. Once mapped, the genome sequence around the locus can be used to design additional genetic markers for fine-scale mapping, and to identify putative candidate genes using the genome annotation. Such genes can be resequenced from informative plants showing phenotypic variation for the target trait. This ability to move directly from "map to genome to gene" will hasten the identification of genes responsible for traits. However, the automated annotation still includes many genes of "unknown function" and there are likely to be as yet unannotated genes in the genome sequence. Moreover, the DM genome represents only one haplotype in a species known to exhibit abundant sequence diversity.

The conversion of ~93% of the assembled genome sequence to well-structured, oriented and annotated PMs has made potato more amenable to modern genomic/genotyping approaches, such as genotyping-by-sequencing (Uitdewilligen *et al.* 2013). The clear and irreversible shift toward sequence based polymorphism in place of 'fragment based' markers will have the effect of augmenting centimorgan positions with genome sequence co-ordinates, providing a means for verifying the accuracy of mapping studies. The integrated DMDD map complements the published potato genome sequence and adds to a growing number of resources for genetic and genomic analyses.

The integrated map presented here and associated resources will help to alleviate many of the complicating aspects of potato as a genetic system. Potato is the most economically important crop where cultivars are highly heterozygous polyploids that suffer severe inbreeding depression on self-pollination. Such breeding systems make breeding and genetical studies difficult and cultivar development generally requires simultaneous recurrent selection for several traits over many years of evaluation. Introduction of traits that would make such crops more sustainable, *e.g.*, drought and salinity tolerance as well as nutrient use efficiency, will be targeted as we confront global climate change and dwindling natural resources (Levy *et al.* 2013). Moreover, attempts to convert the cross-pollinated tetraploid breeding system into an F_1 hybrid diploid based scheme are also in progress (Lindhout *et al.* 2011). The isolation of genes coding for key traits, and characterization of their functional allelic diversity will be greatly facilitated by the resources provided in this study. A recent example is the identification of a gene largely responsible for the adaptation of Andean-derived potato germplasm to the longer day-lengths of temperate latitudes (Kloosterman *et al.* 2013).

The work presented here has generated a greatly improved ordering of the potato reference genome superscaffolds into chromosomal PMs. The reconfigured PMs and their links with genetic maps provide a major new resource for the research community. They form the basis by which geneticists can identify genes underlying important traits and through which comparative genomics can be further exploited in diversity assessment, phylogenetic inference, and plant breeding.

ACKNOWLEDGMENTS

We thank Andrzej Kilian (Diversity Arrays Technology, Australia) for DArT genotyping of the DMDD mapping population. We acknowledge Peter E. Hedley and Clare Booth (The James Hutton Institute, UK) for help with SNP genotyping. We thank S. B. Divito (Instituto Nacional de Tecnología Agropecuaria, Balcarce, Argentina) for technical assistance. We are also grateful to Luke Ramsay and Peter E. Hedley (The James Hutton Institute, UK) for comments on the manuscript. AFLP and WGP are (registered) trademarks owned by KeyGene N.V. We acknowledge the funding made available by the Potato Genome Sequencing grant, UK [Scottish Government Rural and Environmental Science and Analytical Services Division (RESAS), Department for Environment, Food

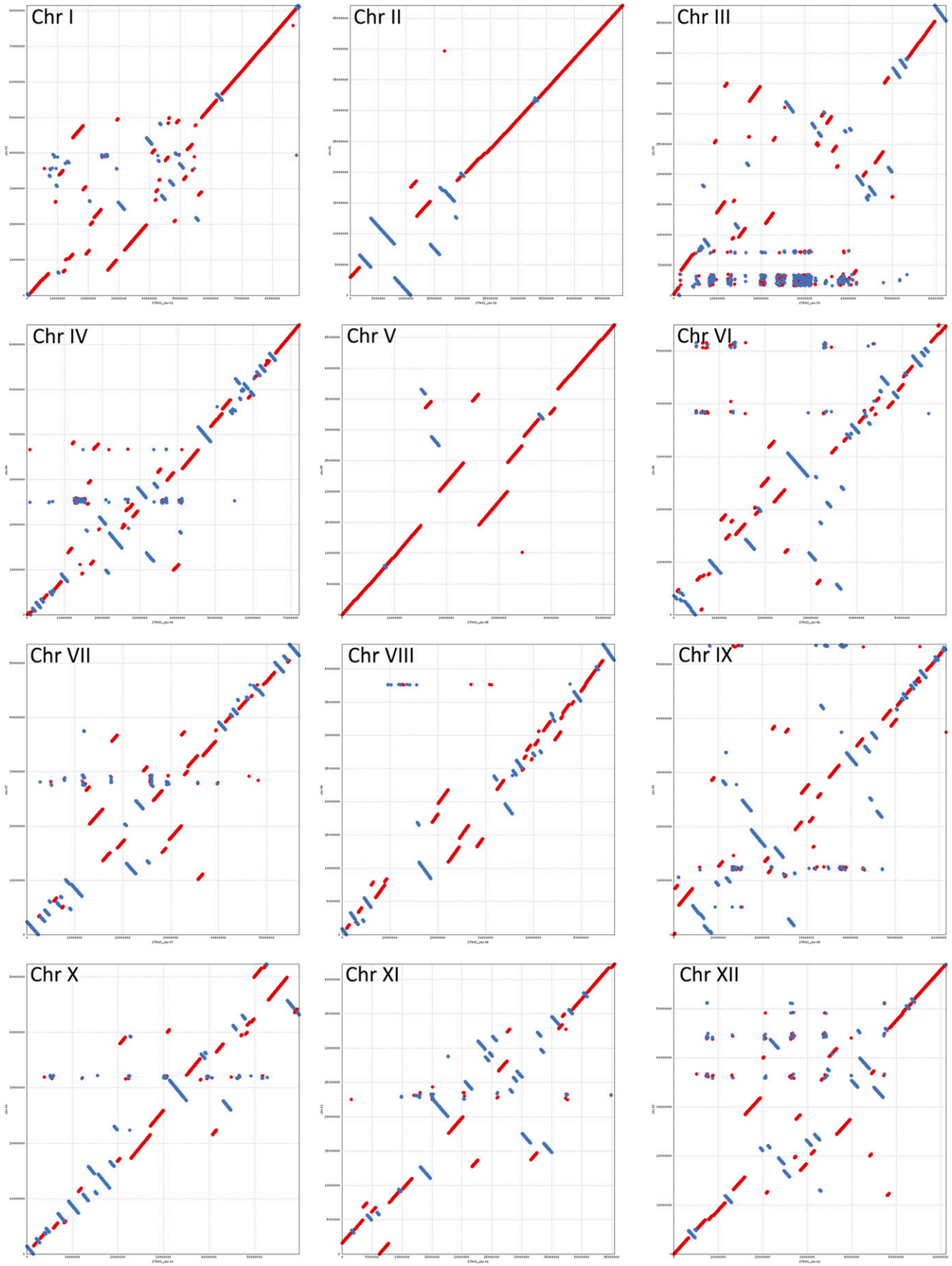

Figure 8 NUCmer sequence alignment dot plots for the twelve potato chromosomes using current (ver4.03, plotted on x-axis) and previous (ver2.1.11, plotted on y-axis) versions of DM PMs. Sequences aligned in forward and reverse orientations are represented by red and blue lines, respectively. Scaffold misplacements are shown as horizontal or vertical shifts in parts of the aligned blocks.

and Rural Affairs (DEFRA), Agriculture and Horticulture Development Board (AHDB)-Potato Council, Biotechnology and Biological Sciences Research Council (BBSRC, Grant BB/F012640)]; New Zealand Institute for Crop & Food Research Ltd Strategic Science Initiative and the New Zealand Institute for Plant & Food Research Ltd Capability Fund, New Zealand; NMEA (Netherlands Ministry of Economic Affairs), CBSG (Centre for BioSystems Genomics), STW (Netherlands Technology Foundation grant 07796), The Netherlands; Teagasc Core Funding, DAFF-Research Stimulus Fund, Ireland; International Potato Center (CIP-CGIAR)/CRP RTB, Peru; CONICYT (Fondap 1509007, Basal CMM, PBCT-PSD-03), CIRIC INRIA, INIA-Ministry of Agriculture of Chile, Chile; FEMCIDI OEA, PE/09/02 MINCyT-CONCyTEC, 2010-2011, Instituto Nacional de Tecnología Agropecuaria (INTA-Core Funds) and Ministerio de Ciencia y Tecnología (MINCyT), Argentina; Proyecto FEMCIDI-OEA SEDI/AE- 305 /09 (2008-2012), Proyecto Bilateral Argentina, Perú; FINCyT (099-FINCyT-EQUIP-2009) / (076-FINCyT-PIN-2008), Prestamo BID no. 1663/OC-PE, Instituto Nacional de Innovacion Agraria, Ministry of Agriculture of Peru, Peruvian Ministry of Agriculture, Technical Secretariat of coordination with the CGIAR, Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica, Peru (CONCYTEC), Special Multilateral Fund of the Inter-American Council for Integral Development (FEMCIDI-Peru).

LITERATURE CITED

Buntjer, J. B., 1999 *Cross Checker*, Vol. 291. Department of Plant Breeding, Wageningen University and Research Centre, Wageningen.

Creste, S., A. T. Neto, and A. Figueira, 2001 Detection of single sequence repeat polymorphisms in denaturing polyacrylamide sequencing gels by silver staining. Plant Mol. Biol. Rep. 19: 299–306.

de Boer, J. M., T. J. A. Borm, T. Jesse, B. Brugmans, L. Wiggers-Perebolte *et al.*, 2012 A hybrid BAC physical map of potato: a framework for sequencing a heterozygous genome (vol 12, 594, 2011). BMC Genomics 13: 423.

Fan, J. B., A. Oliphant, R. Shen, B. G. Kermani, F. Garcia *et al.*, 2003 Highly parallel SNP genotyping. Cold Spring Harb. Symp. Quant. Biol. 68: 69–78.

Feingold, S., J. Lloyd, N. Norero, M. Bonierbale, and J. Lorenzen, 2005 Mapping and characterization of new EST-derived microsatellites for potato (*Solanum tuberosum* L.). Theor. Appl. Genet. 111: 456–466.

Felcher, K. J., J. J. Coombs, A. N. Massa, C. N. Hansey, J. P. Hamilton *et al.*, 2012 Integration of two diploid potato linkage maps with the potato genome sequence. PLoS ONE 7: e36347.

Fulton, T. M., R. Van der Hoeven, N. T. Eannetta, and S. D. Tanksley, 2002 Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14: 1457–1467.

Gebhardt, C., E. Ritter, A. Barone, T. Debener, B. Walkemeier *et al.*, 1991 RFLP maps of potato and their alignment with the homoeologous tomato genome. Theor. Appl. Genet. 83: 49–57.

Ghislain, M., J. Núñez, M. R. Herrera, J. Pignataro, F. Guzman *et al.*, 2009 Robust and highly informative microsatellite-based genetic identity kit for potato. Mol. Breed. 23: 377–388.

Gong, Z., Y. Wu, A. Koblízková, G. A. Torres, K. Wang *et al.*, 2012 Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell 24: 3559–3574.

Green, E. D., and P. Green, 1991 Sequence-tagged site (STS) content mapping of human chromosomes: theoretical considerations and early experiences. PCR Methods Appl. 1: 77–90.

Hamilton, J. P., and C. R. Buell, 2012 Advances in plant genome sequencing. Plant J. 70: 177–190.

Hamilton, J. P., C. N. Hansey, B. R. Whitty, K. Stoffel, A. N. Massa *et al.*, 2011 Single nucleotide polymorphism discovery in elite North American potato germplasm. BMC Genomics 12: 302.

Herrera, M. R., and M. Ghislain, 2000 *Molecular Biology Laboratory Protocols: Plant Genotyping*, Ed. 3. Crop Improvement and Genetic Resources Department, International Potato Center (CIP), Lima, Peru.

International Rice Genome Sequencing Project, 2005 The map-based sequence of the rice genome. Nature 436: 793–800.

Iovene, M., S. M. Wielgus, P. W. Simon, C. R. Buell, and J. M. Jiang, 2008 Chromatin structure and physical mapping of chromosome 6 of potato and comparative analyses with tomato. Genetics 180: 1307–1317.

Istrail, S., G. G. Sutton, L. Florea, A. L. Halpern, C. M. Mobarry *et al.*, 2004 Whole-genome shotgun assembly and comparison of human genome assemblies. Proc. Natl. Acad. Sci. USA 101: 1916–1921.

Jaccoud, D., K. Peng, D. Feinstein, and A. Kilian, 2001 Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res. 29: E25.

Kloosterman, B., J. A. Abelenda, M. M. C. Gomez, M. Oortwijn, J. M. de Boer *et al.*, 2013 Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495: 246–250.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12.

Levy, D., W. K. Coleman, and R. E. Veilleux, 2013 Adaptation of potato to water shortage: irrigation management and enhancement of tolerance to drought and salinity. Am. J. Potato Res. 90: 186–206.

Lindhout, P., D. Meijer, T. Schotte, R. C. B. Hutten, R. G. F. Visser *et al.*, 2011 Towards F 1 hybrid seed potato breeding. Potato Res. 54: 301–312.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader *et al.*, 2006 Genome sequencing in microfabricated high-density picolitre reactors (vol 437, pg 376, 2005). Nature 441: 120–120.

Milbourne, D., R. C. Meyer, A. J. Collins, L. D. Ramsay, C. Gebhardt *et al.*, 1998 Isolation, characterisation and mapping of simple sequence repeat loci in potato. Mol. Gen. Genet. 259: 233–245.

Miller, J. R., S. Koren, and G. Sutton, 2010 Assembly algorithms for next-generation sequencing data. Genomics 95: 315–327.

Ning, Z. M., A. J. Cox, and J. C. Mullikin, 2001 SSAHA: a fast search method for large DNA databases. Genome Res. 11: 1725–1729.

Ovchinnikova, A., E. Krylova, T. Gavrilenko, T. Smekalova, M. Zhuk *et al.*, 2011 Taxonomy of cultivated potatoes (*Solanum* section *Petota*: Solanaceae). Bot. J. Linn. Soc. 165: 107–155.

Park, T. H., J. B. Kim, R. C. Hutten, H. J. van Eck, E. Jacobsen *et al.*, 2007 Genetic positioning of centromeres using half-tetrad analysis in a 4x-2x cross population of potato. Genetics 176: 85–94.

Paz, M. M., and R. E. Veilleux, 1999 Influence of culture medium and in vitro conditions on shoot regeneration in *Solanum phureja* monoploids and fertility of regenerated doubled monoploids. Plant Breed. 118: 53–57.

Peters, S. A., J. W. Bargsten, D. Szinay, J. van de Belt, R. G. Visser *et al.*, 2012 Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. Plant J. 71: 602–614.

Potato Genome Sequencing Consortium, 2011 Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195.

Rezvoy, C. M., D. Charif, L. Gue'guen, and G. A. B. Marais, 2007 MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics 23: 2188–2189.

Sliwka, J., H. Jakuczun, M. Chmielarz, A. Hara-Skrzypiec, I. Tomczyn'ska *et al.*, 2012 A resistance gene against potato late blight originating from Solanum x michoacanum maps to potato chromosome VII. Theor. Appl. Genet. 124: 397–406.

Spooner, D. M., J. Núñez, G. Trujillo, M. D. Herrera, F. Guzmán *et al.*, 2007 Extensive simple sequence repeat genotyping of potato landraces supports a major reevaluation of their gene pool structure and classification. Proc. Natl. Acad. Sci. USA 104: 19398–19403.

Tang, J. F., S. J. Baldwin, J. M. E. Jacobs, C. G. van der Linden, R. E. Voorrips *et al.*, 2008a Large-scale identification of polymorphic microsatellites using an in silico approach. BMC Bioinformatics 9: 374.

Tang, X., D. Szinay, C. Lang, M. S. Ramanna, E. A. van der Vossen *et al.*, 2008b Cross-species bacterial artificial chromosome-fluorescence *in situ* hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. Genetics 180: 1319–1328.

Tang, X., J. M. de Boer, H. J. van Eck, C. Bachem, R. G. Visser *et al.*, 2009 Assignment of genetic linkage maps to diploid *Solanum tuberosum* pachytene chromosomes by BAC-FISH technology. Chromosome Res. 17: 899–915.

Tanksley, S. D., M. W. Ganal, J. P. Prince, M. C. Devicente, M. W. Bonierbale *et al.*, 1992 High-density molecular linkage maps of the tomato and potato genomes. Genetics 132: 1141–1160.

The Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

The French-Italian Public Consortium for Grapevine Genome Characterization, 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463–467.

The Tomato Genome Sequencing Consortium, 2012 The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635–641.

Uitdewilligen, J. G., A. M. Wolters, B. B. D'hoop, T. J. Borm, R. G. Visser *et al.*, 2013 A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE 8: e62355.

Van Ooijen, J. W., 2006 *Joinmap 4: Software for the Calculation of Genetic Linkage Maps*. Kyazma B. V., Wageningen, The Netherlands.

van Os, H., S. Andrzejewski, E. Bakker, I. Barrena, G. J. Bryan *et al.*, 2006 Construction of a 10,000-marker ultradense genetic recombination map of potato: Providing a framework for accelerated gene isolation and a genomewide physical map. Genetics 173: 1075–1087.

Veilleux, R. E., L. Y. Shen, and M. M. Paz, 1995 Analysis of the genetic composition of anther-derived potato by randomly amplified polymorphic DNA and simple sequence repeats. Genome 38: 1153–1162.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Vandelee *et al.*, 1995 AFLP: a new technique for DNA-fingerprinting. Nucleic Acids Res. 23: 4407–4414.

Wenzl, P., J. Carling, D. Kudrna, D. Jaccoud, E. Huttner *et al.*, 2004 Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc. Natl. Acad. Sci. USA 101: 9915–9920.

Communicating editor: D. Zamir

Tree Genetics & Genomes (2014) 10:1353–1367
DOI 10.1007/s11295-014-0765-6

ORIGINAL PAPER

Transcriptome of tung tree mature seeds with an emphasis on lipid metabolism genes

Vanessa Galli · Frank Guzman · Rafael S. Messias · Ana P. Körbes · Sérgio D. A. Silva · Márcia Margis-Pinheiro · Rogério Margis

Received: 27 January 2014 / Revised: 10 June 2014 / Accepted: 16 June 2014 / Published online: 27 June 2014

Abstract Tung oil, the major product of tung tree (*Vernicia fordii*) seeds, is one of the highest quality oils for industrial applications and has been considered for the production of biodiesel. Considering the poor agronomical traits of this crop, efforts have been made to breed tung trees for a higher fruit yield and oil property modification for biodiesel use or to engineer plants to produce a higher tung oil yield. However, these efforts have been hampered by a lack of molecular information, as there is no available genome and identified and characterized transcripts of this tree are scarce. Furthermore, there are still many knowledge gaps regarding tung oil biosynthesis. To provide a comprehensive and accurate foundation for molecular studies of tung tree, we present here a reference transcriptome dataset of mature tung seeds. A set of 43,081,927 reads were assembled into 47,585 unigenes. A homology search using blastx against the GenBank nonredundant protein database and the Swiss-Prot database resulted in the annotation of 96 and 81 % of these unigenes, respectively. We also systematically arranged a series of unigenes potentially associated with oil biosynthesis and degradation and examined the expression profile of a subset of those genes in samples from five different stages of seed development, providing a valuable source of genes and transcriptional information related to these pathways. This study represents the first large-scale transcriptome annotation of tung tree and will be useful in tung breeding for oil properties and other agronomical traits.

Keywords *Vernicia fordii* · RNA-Seq · Seed development · Oil synthesis · Oil breakdown

Communicated by J. F. D. Dean

Electronic supplementary material The online version of this article (doi:10.1007/s11295-014-0765-6) contains supplementary material, which is available to authorized users.

V. Galli · R. Margis (✉)
Centro de Biotecnologia, PPGBCM, Laboratório de Genomas e População de Plantas, prédio 43431, Universidade Federal do Rio Grande do Sul-UFRGS, P.O. Box 15005, CEP 91501-970 Porto Alegre, Rio Grande do Sul, Brazil
e-mail: rogerio.margis@ufrgs.br

V. Galli · R. S. Messias · S. D. A. Silva
Empresa Brasileira de Pesquisa Agropecuária-EMBRAPA, P.O. Box 403, CEP 96010-971 Pelotas, Rio de Grande do Sul, Brazil

F. Guzman · A. P. Körbes · M. Margis-Pinheiro · R. Margis
PPGGBM, Universidade Federal do Rio Grande do Sul-UFRGS, P.O. Box 15005, CEP 91501-970 Porto Alegre, Rio Grande do Sul, Brazil

Introduction

As a result of our growing population, diminishing petrochemical resources, and environmental consciousness, there will be a worldwide increasing demand of any renewable energy supply that does not cause adverse environmental impacts and does not compete with the food supply. Crop plants offer a substantial potential to provide renewable chemical feedstock that could alleviate the use of petroleum-based products in industrial applications (Dyer and Mullen 2005; Vanhercke et al. 2013). Tung oil, the major product of tung tree (*Vernicia fordii*) seeds, is considered one of the highest quality oils. Tung seeds accumulate high levels of α-eleostearic acid (approximately 80 %), a trienoic fatty acid with conjugated double bonds (9*cis*, 11*trans*, and 13*trans* octadecatrienoic acid), which is widely used in paints, high-quality printing, plasticizers, medicine, and chemical reagents. Moreover, because tung seeds accumulate a high content of oil (approximately 50 %), this species has recently been considered for use in biodiesel production. However, the large-scale production of tung oil through traditional farming is hampered because of the poor agronomic traits of this plant

species (Park et al. 2008; Shang et al. 2010). Therefore, increasing the yield and adjusting the characteristics of tung oil are major challenges for industry. Several breeding approaches, from traditional breeding to DNA marker-assisted selection and genetic engineering, are being deployed to meet these objectives (Brown and Keeler 2005).

There is also interest in the manipulation of lipid metabolism in conventional, easier-to-grow oilseed crops, such as soybean (*Glycine max*) and palm (*Butia capitata*), to produce a tung-specific drying oil. Unfortunately, the expression of certain transgenes has shown limited success due to the low accumulation of the desired fatty acids in the transgenic plants (Jaworski and Cahoon 2003; Cahoon et al. 2006; Rupilinus and Ahmad 2007; Vanhercke et al. 2013). It is clear from these studies that significantly more knowledge of the synthesis, accumulation, and storage of plant oils is needed to efficiently synthesize and accumulate the unusual fatty acid from tung oil in transgenic hosts.

Over the past several years, we have greatly improved our understanding of a plethora of biological processes, including lipid metabolism, through the sequencing of a large number of genomes and transcriptomes from several plant species (Meier et al. 2011; Natarajan and Parani 2011; Schmucki et al. 2013). Currently, the lack of an available *V. fordii* genome prevents the development of reliable genome-based tools to evaluate gene expression dynamics by microarrays or qPCR panels and the study of genes for metabolic engineering approaches. Furthermore, only 2,851 sequences of *V. fordii* are available in the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov, accessed in January 2014). Therefore, efforts to identify novel genes in this plant are of great interest. Transcriptome sequencing is an efficient methodology for large-scale gene discovery, and next-generation sequencing technologies, such as RNA-Seq, have emerged as cost-effective and massive unbiased approaches to sequencing complementary DNAs (cDNAs) without cloning. Moreover, the large number of short reads and depth of coverage generated by the RNA-Seq approach enables the de novo assembly of sequences for the construction of the transcriptome of nonmodel organisms without genome sequences. Such a transcriptome can be exploited for gene annotation and discovery, molecular marker development, genomic and transcriptomic assembly, and microarray development; it also provides a global measurement of transcript abundance (Crawford et al. 2010; Wang et al. 2010; Venturini et al. 2013).

To provide a comprehensive and accurate foundation for molecular studies of tung tree, we present here a reference transcriptome dataset from mature seeds of *V. fordii*, assembled and annotated from deep RNA-Seq data. This study represents the first large-scale transcriptome annotation of tung tree. We also systematically arranged a series of transcripts potentially associated with lipid metabolism and examined the expression profile of a subset of those genes in samples from different stages of seed development. This approach provided a valuable source of genes involved in seed oil biosynthesis and other agronomic traits that will be useful for engineering other plants to produce tung oil or for breeding tung tree for a higher fruit yield and for modified oil properties for use as biodiesel.

Methods

Plant material and RNA isolation

For the construction of the mRNA-Seq library, fruits from *V. fordii* plants grown in an open environment at Embrapa Clima Temperado (Pelotas, Brazil) were collected at mature stage S6 (approximately 120 days after flower opening (DAF)); fruits from the S1 (20 DAF), S2 (35 DAF), S3 (50 DAF), S4 (80 DAF), and S5 (100 DAF) stages were collected to perform the expression analysis. Figure S1 provides a visual representation of each of the six stages. Seeds were dissected from all collected fruits, pooled, and immediately frozen in liquid nitrogen and stored at −80 °C. Total RNA was isolated using Trizol reagent (Invitrogen, CA, USA) according to the manufacturer's protocol. The RNA quality was evaluated by electrophoresis through a 1 % agarose gel, and the RNA concentration was determined by absorbance at 260 nm using a Nanodrop spectrophotometer (Nanodrop Technologies, USA).

RNA-Seq library construction using deep sequencing

Total RNA (>10 μg) isolated from mature seeds (120 DAF) was sent to Fasteris SA (Plan-les-Ouates, Switzerland) for processing and sequencing using an Illumina HiSeq2000. To generate the RNA-Seq library, polyadenylated transcript sequencing was performed as follows: poly-A mRNA was purified and cDNA was synthesized using poly(T) primer and a shotgun approach to generate inserts of approximately 500 nt. The 3P and 5P adapters were bound, and a cDNA colony template library was generated by PCR amplification and single-end sequenced with Illumina.

Data filtering and de novo assembly

The overall chart flow for analyzing the Illumina RNA-Seq library is shown in Fig. 1. The initial base setup and quality filtering of the image data were performed using the default parameters in the Illumina data processing pipeline. Thereafter, all low-quality reads (FASTq value of <13) were removed, and the 5P and 3P adapter sequences were trimmed using Genome Analyzer Pipeline (Fasteris SA). The remaining low-quality reads with undetermined nucleotides (nt) were

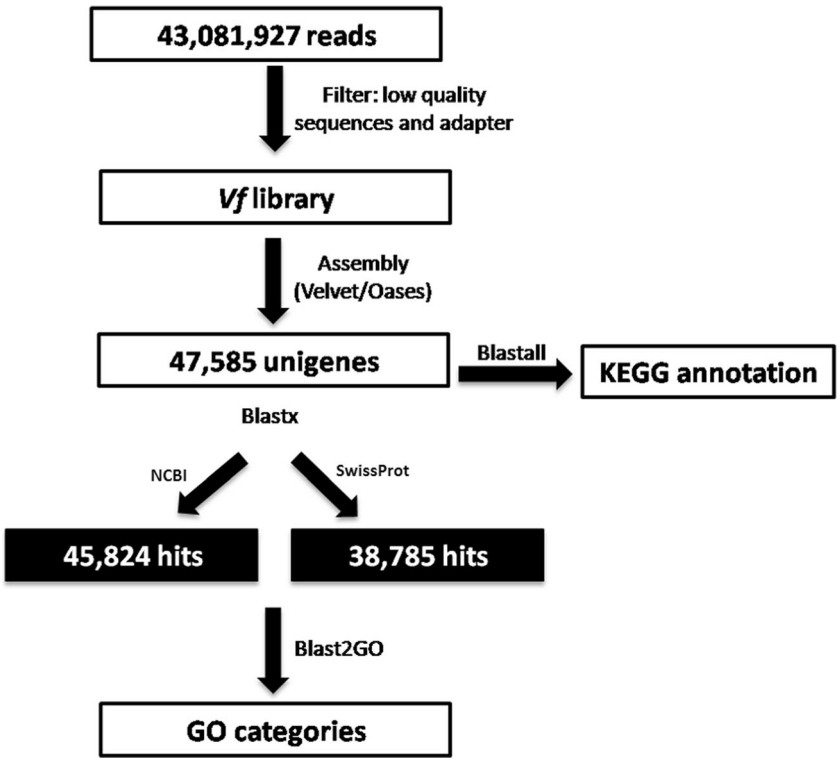

Fig. 1 Bioinformatics pipeline for the assembly of reads and annotation of contigs. We generated 43,081,927 sequences of reads from tung mature seeds (120 days after flower opening). After quality filtering, the reads were assembled into 47,585 contigs that were aligned against the GenBank nonredundant protein database (NR) and Swiss-Prot database, resulting in 45,824 and 38,785 matches, respectively. Gene ontology (*GO*) terms were obtained from Blast2Go software, and the KEGG pathway annotations were performed using Blastall software against the KEGG database

removed using PrinSeq script (Schmieder and Edwards 2011). After data cleaning (low-quality reads, adapter sequences), the RNA-Seq data were de novo assembled into contigs using the Velvet/Oases package (Schulz et al. 2012). We used a minimum contig length of 200, and a multi-*k*-mer (i.e., 21, 31, 41, 51, and 61 bp; substrings of length *k*)-based strategy to capture the most diverse assembly with improved specificity and sensitivity, especially for low-expressed genes. We then used the USEARCH algorithm (Edgar 2010) to obtain unigenes. The metrics used to assess the transcriptome assembly quality included the overall number (coverage) of contigs, the average length of contigs, and the diversity of contigs (the estimated number of nonredundant (NR) contigs). The result of this analysis corresponds to the final number of unigenes or independent tung transcripts. To determine transcript abundances, high-quality reads were mapped to the assembled transcriptome using Bowtie software (v0.12.7) (Langmead et al. 2009). Reads mapping to each unigene were counted using SAM tools (v0.1.16).

Gene annotation and analysis

All unigenes were utilized for homology searches against protein databases such as NR sequences from NCBI (http://www.ncbi.nlm.nih.gov/) and the Swiss-Prot database (http://www.expasy.ch/sprot/) by applying the BLASTX program (*e* value $<1e^{-6}$); the best-aligning results were selected to annotate the unigenes. If the aligning results from the databases were in conflict with each other, the results from the Swiss-Prot database were preferentially selected.

The Blast2GO program (Conesa and Götz 2008) was used to assign putative functionalities, GO terms, and Kyoto Encyclopedia of Genes and Genomes (KEGG)-based metabolic pathways. Final GO assignments were defined based on level 2. All other settings for the analysis were maintained as the defaults. The WEGO software was then used to perform GO functional classification of all unigenes to view the distribution of gene functions within different pathways at the macro level (Ye et al. 2006). This analysis mapped all of the annotated unigenes to GO terms in the database and calculated the number of unigenes associated with every GO term. KEGG pathway annotations were performed using Blastall software against the KEGG databases (http://www.genome.jp/kegg/). A graphic representation of assigned metabolic pathways was obtained using the IPATH2 software (http://pathways.embl.de/ipath2). The unigenes with a functional assignment as a fatty acid desaturase were translated into peptides by querying the longest predicted open reading frame (ORF) using the ORF Finder tool (http://www.ncbi.nlm.nih.gov/gorf/gorf.html).

Expression profile of genes involved in lipid metabolism during seed development

RT-qPCR was performed to evaluate the expression profile of selected genes involved in lipid metabolism during tung seed development. Three replicates of RNA samples from the seeds of tung fruits at different developmental stages, corresponding to 20 (S1), 35 (S2), 50 (S3), 80 (S4), and 100 (S5) DAF, as presented in Fig. S1, were extracted using the Trizol reagent

(Invitrogen, CA, USA). The RNA quality was assessed by 1 % agarose gel electrophoresis. Total RNA (1 μg) was digested with 1 U DNAse I and DNAse 1× reaction buffer and reverse transcribed using the M-MLV enzyme and oligo-24TV primers according to the manufacturer's instructions (Invitrogen). Specific primers for amplification of the selected genes were designed using Vector NTI10 software (Invitrogen, CA, USA) (Table S1). A melting temperature of 58–62 °C, a GC content of 45–55 %, and an amplicon size below 250 bp was defined for all primer pairs. The specificity of the amplicons was verified by the presence of a single peak for RT-qPCR melting curve products and a single band of the expected size on a 3 % agarose gel (data not shown). The cDNAs were amplified by RT-qPCR in a final volume of 20 μL containing 1 μL cDNA, 10 μL of Platinum Sybr green UDG (Invitrogen), and 2–5 pmol of each primer. The amplification was standardized using a 7500 Real time Fast thermocycler (Applied Biosystems) with the following conditions: 50 °C for 20 s, 95 °C for 10 min, followed by 45 cycles of 15 s at 95 °C and 60 s at 60 °C. The PCR products for each primer set were subjected to a melting curve analysis to verify the presence of primer dimers or nonspecific amplicons. The melting curve analysis ranged from 60 to 95 °C, increasing the temperature stepwise by 1 %. No-template controls and a reverse transcription negative control were included to ensure no reagent or genomic DNA contamination. Genes coding for *actin*, *ubiquitin*, and *tubulin* were used as internal controls. The relative expression data were calculated according to the $2^{-\Delta\Delta Cq}$ method and are presented as fold change (Livak and Schmittgen 2001). Samples from the S1 stage were used as calibrator samples. Statistical analyses were performed using the computer program SAS System for Windows v9.1.3. The data were subjected to a variance analysis ($p \leq 0.05$). In the case of statistical significance, the relative expression among the stages of seed development was compared by a Tukey test ($p \leq 0.05$) and denoted by the use of different letters.

Results and discussion

Tung seed expression sequence database

We sequenced 43,081,927 reads from mature tung seeds using Illumina sequencing. For analysis, the contigs were divided into 21, 31, 41, 51, and 61 *k*-mers to improve the specificity and sensitivity of the assembly using Velvet. Therefore, 97,647, 82,023, 69,855, 62,041, and 51,157 transcripts were obtained using 21, 31, 41, 51, and 61 *k*-mers, respectively, for assembly. The mean size of the transcripts was 1,108, 1,223, 1,276, 1,191, and 1,266 bp for the 21, 31, 41, 51, and 61 *k*-mers, respectively. The statistics of the *V. fordii* transcripts obtained with the different *k*-mers using Velvet is shown in Table 1. The use of multi *k*-mers for *de novo* assembly has been successfully implemented by several authors (Surget-Groba and Montoya-Burgos 2010; Garg et al. 2011; Gruenheit et al. 2012; Yang et al. 2012). All contigs were further merged by integrating sequence overlaps to determine the number of unique sequences. Therefore, 47,585 unisequences (unigenes) were obtained, creating an initial reference transcriptome. The lengths of the unigenes ranged from 200 to 19,718 nt, with an average size of 1,684 nt. From the unigenes obtained, 64 % were more than 1,000 nt, confirming the quality of the transcriptome assembled. The distribution of the unigenes according to length is presented in Fig. 2, with most of the unigenes being between 1,000 and 2,500 nt.

All unigenes were aligned against the NR protein database of GenBank using BLASTX with an *e* value cut-off of $1e^{-6}$. We found matches for 45,824 unigenes (96 %). The best hit from each annotated sequence was calculated and is presented in Table S2. A majority of the best hits were from *Vitis vinifera* (17,784 sequences, 38.8 %), most likely because there are a large number of deposited ESTs of this species in the NR database. The second most frequent species was *G. max* (5,546 sequences, 12.1 %), followed by *Populus trichocarpa* (4,665 sequences, 10.2 %) and *Arabidopsis thaliana* (4,398 sequences, 9.57 %), as shown in Fig. 3a. Only 213 sequences (0.46 %) matched the sequences from *V. fordii* deposited in the NR database, confirming the low number of publicly available sequences for this plant. We also aligned the unigenes against the Swiss-Prot (SW) database, which resulted in the annotation of 38,785 unigenes (81 %). The best hit from each annotated sequence was calculated and is presented in Table S3. As shown in Fig. 3b, *A. thaliana* was the most frequent species in this analysis (20,278 sequences, 52.3 %), most likely because the SW database is enriched with sequences from this species as a plant model. Tables S2 and S3 show the best hit for the alignment of unigenes against the

Table 1 Statistics of *Vernicia fordii* transcripts obtained with different *k*-mers using Velvet

Description	*k*-mer 21	*k*-mer 31	*k*-mer 41	*k*-mer 51	*k*-mer 61	Total
Number of contigs	97,647	82,023	69,855	62,041	51,157	362,723
Median contig length	577	785	887	823	868	3,940
Mean contig length	1,108	1,223	1,276	1,191	1,266	6,064
Max contig length	17,304	19,718	16,856	16,123	16,607	86,608
Number of contigs >1 kbp	37,180	35,871	32,372	27,023	23,442	155,888
N50	2,170	2,187	2,146	1,931	2,092	10,526

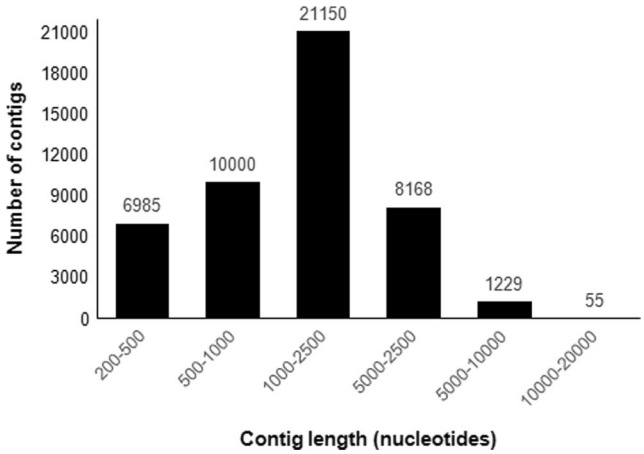

Fig. 2 Size distribution and abundance of contigs from the tung transcriptome

NR and SW databases, respectively, including the bit score and *e* value of the alignment, information regarding the annotated sequence such as nt and protein accession number, the coding sequence (CDS) structure, and species with homologous sequences.

The KEGG pathway annotation for 40,392 of the unigenes was obtained using Blastall software against the KEGG database, as presented in Table S4. Additionally, a graphic representation with the association of each unigene with metabolic pathways was obtained using the IPHATH2 software (Fig. S2). As observed in Table S4 and Fig. S3, the unigenes are widely distributed in distinct metabolic pathways, confirming the large coverage of the transcriptome obtained.

Of note is the presence of several unigenes related to lipid metabolism, such as 126 unigenes classified as lipid biosynthesis-related proteins, 84 belonging to linoleic acid metabolism, 76 related to fatty acid biosynthesis (FAS), 73 related to the biosynthesis of unsaturated fatty acids, 8 from fatty acid elongation in mitochondria, and 136 from fatty acid metabolism. The annotated unigenes were also classified according to Blast2GO categorization (Fig. 4). According to the categorization "cellular component" GO annotation, the majority of the unigenes belong to the plastid, followed by the mitochondria and plasma membrane (23, 17, and 11 %, respectively). In the category "molecular functions," 18 % of the unigenes encode for protein-binding proteins, 14 % encode for proteins with hydrolase activity, and 12 % of the unigenes have catalytic activity. Some of them (1 %) also participate in lipid binding. With respect to biological processes, the *V. fordii* unigenes are involved in a broad range of physiological functions, especially transport (18 %), protein metabolic process (12 %), and cellular protein modification process (10 %), and almost 3 % of them are related to lipid metabolic processes.

Categories of the most abundant contigs in mature tung seeds

Through the alignment of the contigs against the unigenes, we were able to estimate the abundance of each unigene in the mature tung seeds. Among the top 50 most highly expressed contigs in mature seeds were those coding for storage proteins belonging to the glutelin and legumin families and late embryogenesis abundant (LEA) proteins (Table S5). These

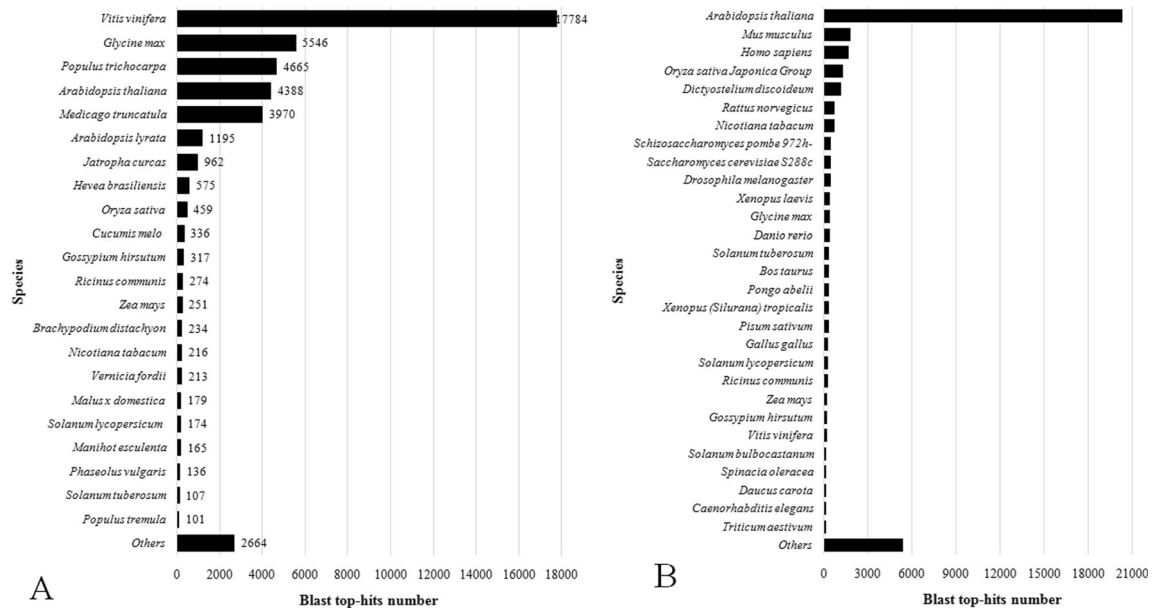

Fig. 3 Blast top-hits resulting from annotation using the GenBank nonredundant protein database (**a**) and Swiss-Prot database (**b**). Contigs of the tung transcriptome were utilized for homology searches against nonredundant protein sequences from NCBI (http://www.ncbi.nlm.nih.gov/) and protein sequences from Swiss-Prot (http://www.expasy.ch/sprot/) by applying the BLASTX program (*e* value $<1e^{-6}$). The frequency of homologous sequences from each species used to annotate the tung sequences was calculated, and the most frequent species are shown

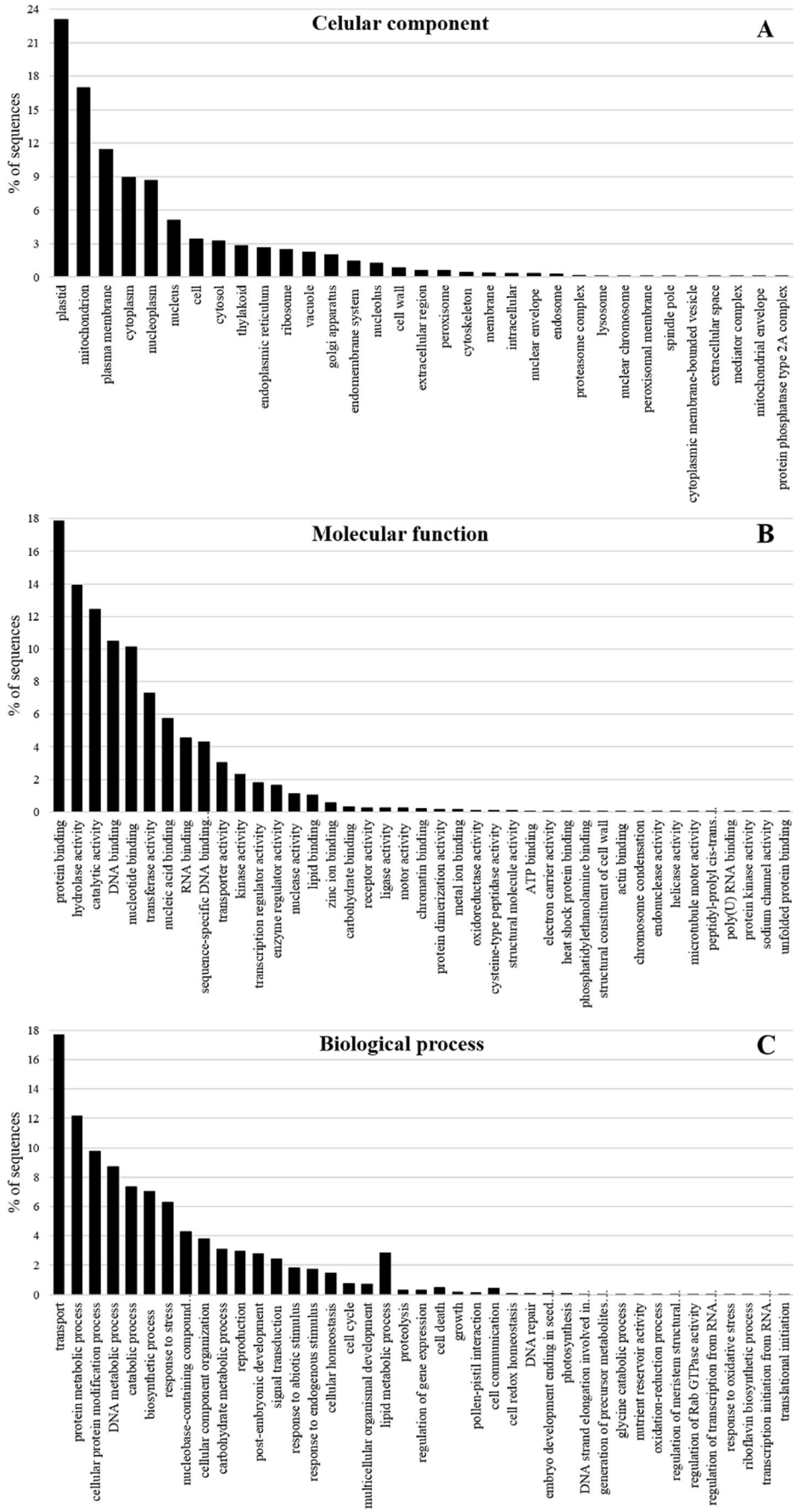

Fig. 4 Functional classification of tung unigenes according to gene ontology (*GO*) terms. The Blast2GO program was used to assign GO terms based on level 2. The percentage of sequences from each GO term was determined

findings are consistent with the high levels of storage proteins in the endosperm of oilseeds (Jolivet et al. 2011). The other most abundant unigenes are related to vacuolar processing and transcriptional regulation. It is interesting to note that among the 100 most abundant unigenes is diacylglycerol *O*-acyltransferase 1 (DGAT1), a well-characterized gene related to triacylglycerol (TAG) biosynthesis (Shockey et al. 2006), as will be discussed further. Additionally, it is important to note that the only gene from lipid metabolism among the most abundantly expressed unigenes was the DGAT1 gene, most likely because we used mature seeds for RNA sequencing because the genes from this pathway typically show a bell-shaped expression pattern during seed development (Chen et al. 2007; Chandran et al. 2014).

Metabolic pathways related to oil accumulation in tung seeds

Tung has been highlighted as a promising target for the production of biodiesel and other petroleum-based industrial products. However, making it an ideal biodiesel crop requires genetic manipulations for increased oil yield and modified oil composition using the genes involved in oil biosynthesis pathway. Therefore, the unigenes that were mapped to the oil biosynthesis pathway based on annotation against the Swiss-Prot and NR databases were analyzed in detail.

In this study, we found 523 NR sequences coding for 382 enzymes and 141 proteins that are involved in oil biosynthesis, encompassing all steps of this pathway, as shown in Fig. 5. The description of these unigenes can be found in Table S6, including information regarding length, abundance, and the species with the best-hit homologous sequence. The expression profile of a subset of these sequences was analyzed by RT-qPCR using samples from five stages of tung seed development (Fig. 6).

Fatty acid biosynthesis

The first committing step in FAS is the conversion of acetyl-CoA to malonyl-CoA, which is catalyzed by acetyl-CoA carboxylase (ACCase) (Fig. 5). There are two types of

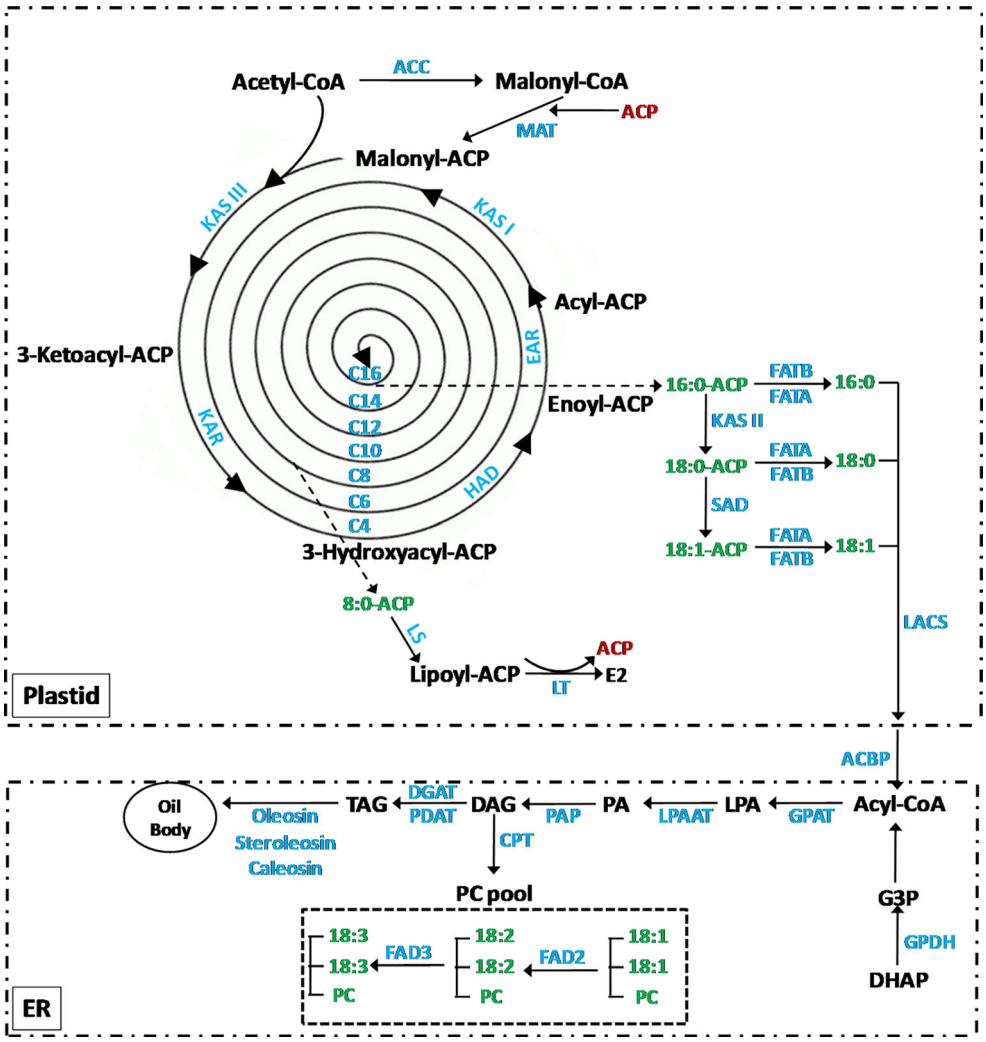

Fig. 5 Schematic representation of the oil biosynthetic pathway. The enzymes for which genes were identified in the tung transcriptome are shown in *light blue*. *ACC* acetyl-CoA carboxylase, *ACP* acyl carrier protein, *MAT* malonyl-CoA acyl transferase, *KAS* 3-ketoacyl-ACP, *KAR* 3-ketoacyl-ACP reductase, *HAD* hydroxyacyl-ACP dehydratase, *EAR* enoyl-ACP reductase, *LS* lipoic acid synthase, *LT* lipoyltransferase, *FATA and FATB* acyl-ACP thioesterases, *SAD* stearoyl ACP desaturase, *LACS* long-chain acyl-CoA synthetase, *ACBP* acyl-CoA-binding protein, *ER* endoplasmic reticulum, *GPAT* glycerol-3-phosphate acyltransferase, *DHAP* dihydroxyacetonephosphate, *GPDH* glycerol-phosphate dehydrogenase, *G3P* glycerol 3-phosphate, *LPA* lysophosphatidic acid, *LPAAT* lysophosphatidic acid acyl transferase, *PA* phosphatidic acid, *PAP* phosphatidic acid phosphatase, *DAG* diacylglycerol, *DGAT* diacylglycerol acyltransferase, *TAG* triacylglycerol, *PDAT* phospholipid:diacylglycerol acyltransferase, *CPT* choline phosphotransferase, *PC* phosphatidylcholine, *FAD2* delta-12-fatty acid desaturase, *FAD3* omega-3-fatty acid desaturase

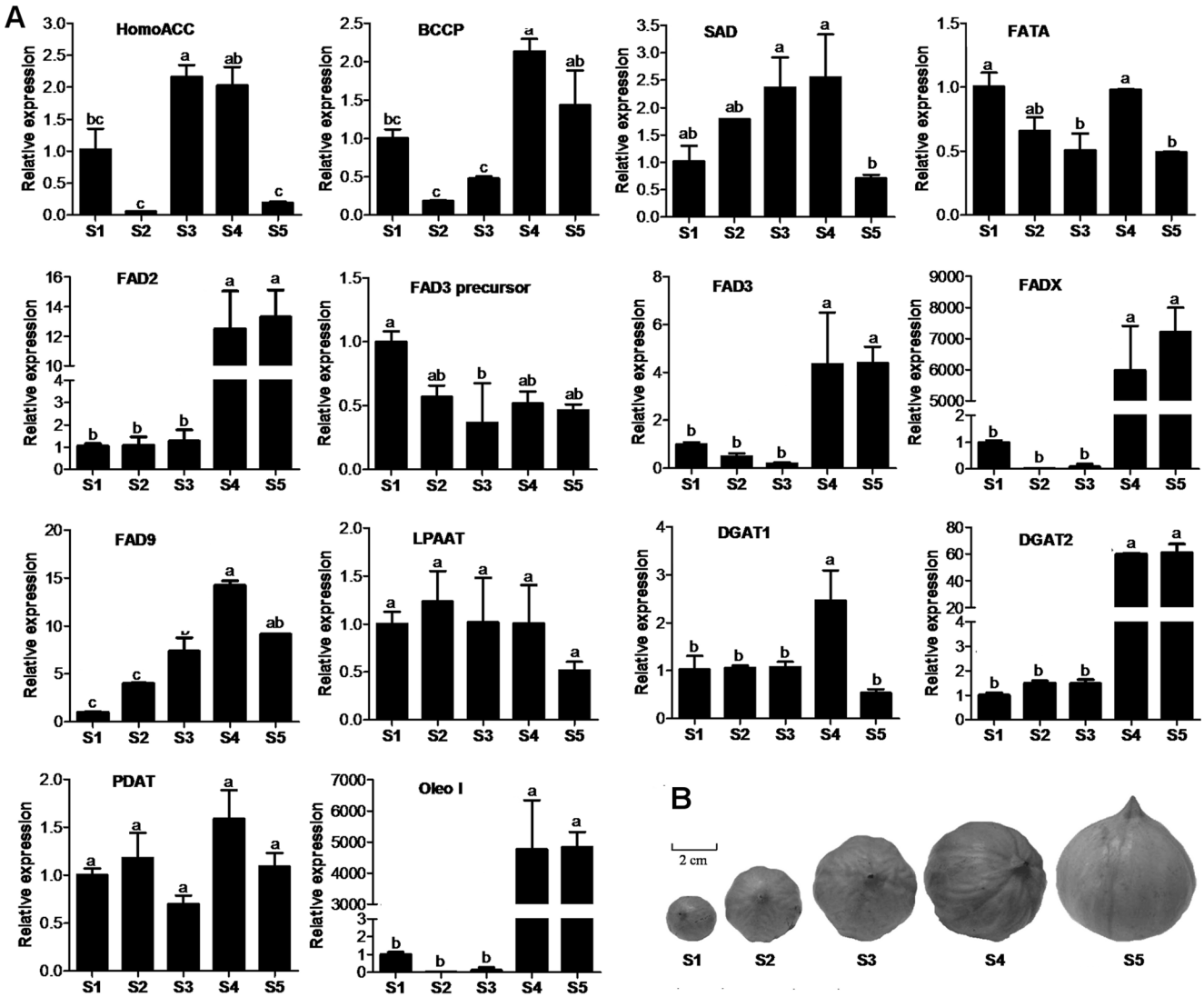

Fig. 6 Expression profile of genes related to lipid accumulation in tung seeds. **a** Expression analyses were performed by RT-qPCR using the S1 stage as a reference sample. *HomoACC* homomeric acyl carboxylase, *BCCP* biotin carboxyl carrier protein, *SAD* stearoyl acyl carrier protein desaturase, *FATA* acyl-ACP thioesterase, *FAD2* delta-12-fatty acid desaturase, *FAD3* omega-3-fatty acid desaturase, *FAD9* delta-9 fatty acid desaturase, *FADX* fatty acid conjugase, *LPAAT* lysophosphatidic acid acyl transferase, *DGAT* diacylglycerol acyltransferase, *PDAT* phospholipid/diacylglycerol acyltransferase, *OleoI* oleosin I. **b** Total RNA was isolated from five stages of seed development

ACCases in higher plants: homomeric and heteromeric ACCase. The carboxylation reaction is catalyzed by heteromeric ACCase, which consists of four subunits that are thought to be coordinately expressed (Ke et al. 2000): a plastid-encoded subunit, carboxyl transferase b-subunit (β-CT), and three nuclear-encoded subunits named biotin carboxyl carrier protein (BCCP), biotin carboxylase (BC), and carboxyl transferase α-subunit (α-CT) (Nakkaew et al. 2008; Li-Beisson et al. 2013). We found 22 unigenes for BCCP (Fig. 5; Table S6). The RT-qPCR analysis from the different stages of tung seed development showed that homomeric ACCase (HomoACC) is more expressed in stages S3 and S4, whereas BCCP is more expressed in stages S4 and S5 (Fig. 6). A similar expression pattern was observed for *Jatropha curcas* (Xu et al. 2011), oil palm (Nakkaew et al. 2008), and castor bean (Chen et al. 2007) seeds, suggesting that both enzymes play an important role in oil accumulation in these seeds. Considering that the expression of ACC genes was also correlated with oil content in oil palm seeds (Nakkaew et al. 2008), the step they catalyze is most likely a rate-limiting step in FAS, and these genes may be useful as molecular markers to assist the selection of high oil-producing varieties or as targets in transgenic approaches aiming to increase the overall oil content. However, Andre et al. (2012) noted ACC as an enzymatic target of feedback inhibition in the FAS pathway, which should be considered during transgenic efforts.

After the ACC step, malonyl-CoA acyl transferase (MAT) catalyzes the condensation of the generated malonyl-CoA and acyl carrier protein (ACP) to form malonyl-ACP. Then, seven

cycles of subsequent condensation, reduction, and dehydration reactions carried out by ketoacyl synthase III and I, 3-ketoacyl-ACP reductase (KAR), hydroxyacyl-ACP dehydratase (HAD), and enoyl-ACP reductase (EAR) (Li-Beisson et al. 2013) are necessary to generate an 18-carbon fatty acid. During this process, some of the octanoic acids generated are converted by lipoic acid synthase (LS) to lipoic acid, the lipoyl group of which is transferred by lipoyltransferase (LT) to apoproteins, such as E2, a component of the plastidial pyruvate dehydrogenase complex (PDHC) that catalyzes the oxidative decarboxylation of pyruvate to acetyl-CoA, thereby beginning a new FAS cycle (Wada et al. 2001). The final elongation of the 16-carbon palmitoyl-ACP to the 18-carbon stearoyl-ACP is catalyzed by KASII (Fig. 5). The ratio of 16-carbon and 18-carbon fatty acids in oil defines its viscosity: the longer the chain length, the higher the viscosity (Allen et al. 1999). As this property affects the atomization of biodiesel, efforts have been made to reduce the viscosity of this oil by lowering the 18-carbon fatty acid content by silencing the activity of KASII (Nguyen and Shanklin 2009). Unigenes of *V. fordii* for all the enzymes and proteins for the FAS step were found in the current study, as shown in Fig. 5. However, the expression profile of these genes in tung tissues remains to be elucidated.

Saturated fatty acids

The synthesis of fatty acids may be accomplished by producing 16:0-ACP fatty acids, which are hydrolyzed by acyl-ACP thioesterases (FATA and FATB) that release fatty acids from the ACP molecule to be transported to the endoplasmic reticulum (ER) (Fig. 5). However, the 18:0-ACP generated by FAS may be desaturated to produce unsaturated fatty acids before being released from ACP and transported to the ER. Desaturases play a pivotal role in fatty acid desaturation during fatty acid and lipid biosynthesis. In plants, 18:0 stearoyl-ACP is desaturated to 18:1 oleoyl-ACP by stearoyl ACP desaturase (SAD). The second and third desaturations, producing 18:2 linoleoyl-ACP and 18:3 linolenyl-ACP, are performed by delta-12-fatty acid desaturase (FAD2) and omega-3-fatty acid desaturase (FAD3), respectively. In the present study, we found unigenes corresponding to SAD, FAD2, FAD3, FATA, and FATB (Fig. 5). The RT-qPCR results (Fig. 6) showed that tung SAD has a bell-shaped expression pattern, with the highest expression at the S2, S3, and S4 stages of seed development (Fig. 6). A similar expression pattern was observed for other oilseed crops with a high content of unsaturated and conjugated fatty acids, such as physic nuts (Jiang et al. 2012) and castor bean (Chen et al. 2007). Considering that only one copy of SAD was found in the tung transcriptome, it is most likely that this SAD copy is responsible for the production of high levels of C18:1 fatty acids at the late–middle stages of seed development, components that will be used in the late stages as sources for the production of α-eleostearic acids. In accordance, the expression of FAD2 and FAD3 was greatly increased at the end of seed development (Fig. 6), concurrent with the high content of polyunsaturated fatty acids in tung seeds produced from the C18:1 fatty acids generated earlier by SAD. Therefore, the transcriptional expression of these genes is closely correlated with TAG accumulation within developing tung seeds. The increased expression of FAD2 and FAD3 at the later stages of seed development, corresponding to seed maturation stages previously described in several oilseed crops such as flax (Banik et al. 2011), linseed (Rajwade et al. 2014), and brassica (Hu et al. 2009), confirm the importance of these enzymes in oil desaturation.

FATA efficiently catalyzes the removal of ACP from 18:1-ACP and has a lower activity toward 18:0- and 16:0-ACP, whereas FATB is more specific for saturated fatty acyl chains (Jiang et al. 2012). Therefore, we evaluated the expression of tung FATA during seed development and found that this gene is constantly expressed at all stages, with a sharply reduced expression at stages S3 and S5 (Fig. 6). Although the number of reads in the tung transcriptome anchored in the FATA unigenes were higher than those anchoring the FATB unigenes (Table S6), it is possible that FATB may assist FATA in the release of ACP from 18:1-ACP, which shows a higher content than 16:0-ACP and 18:0-ACP in tung seeds. It is also possible that FATA has a housekeeping function, providing the constant demand of fatty acids for membrane lipid biosynthesis, in addition to assembly into TAG. Overall, the fatty acid composition of the oil in the developing tung seeds is consistent with the relative expression levels of the SAD, FATA, FAD2, and FAD3 genes, suggesting that these genes have a great contribution to the high content of polyunsaturated fatty acid in tung seeds (Shang et al. 2010).

In tung seeds, the most common fatty acids (more than 80 %) are conjugated fatty acids, such as α-eleostearic acid ($18{:}3^{9cis,11trans,13trans}$). The typical mechanism for generating conjugated fatty acids in plants involves fatty acid oxidation and bond rearrangement. The enzymes capable of synthesizing conjugated fatty acids are called conjugases and are closely related in terms of their amino acid identity to the FAD2 family. Dyer et al. (2002) showed that a single divergent FAD2 enzyme isolated from tung, named FADX, can act upon each of the common unsaturated fatty acids in plants (oleic, linoleic, and linolenic acids) to produce three different unusual fatty acids ($18{:}2^{\Delta 9cis,12trans}$, α-eleostearic, and α-parinaric acids, respectively). Furthermore, Dyer et al. (2002) showed that, unlike FAD2, which is expressed in leaves and seeds, FADX is only expressed in the seed of tung plants. The full-length sequence of the plant common FAD2 (383 amino acids) and FAD3 (458 amino acids) desaturases and the full-length sequence of FADX (386 amino acids) were identified in the present study (Table 2). According to the RT-

Table 2 Fatty acid desaturases from *Vernicia fordii* identified in the transcriptome

Unigene	Confidence	Length	Reads	ORF structure (nt)	ORF length	ORF size (aa)	Gene description	Abbreviation	Reference	Coverage	Identity
31_Locus_163_contig_5_8	0.222	1,460	25,776	1,052–467	Incomplete	206	Delta 12 oleic acid desaturase	FAD2	*Vernicia fordii*	99	96
41_Locus_186_contig_5_6	0.357	1,965	47,280	913–1,597	Incomplete	224			*V. fordii*	99	99
51_Locus_184_contig_5_6	0.333	1,961	55,745	432–1,583	Full	383			*V. fordii*	99	99
31_Locus_16_contig_5651_6924	1	1,362	4,014	1,337–177	Full	386	Delta 12 fatty acid conjugase	FADX	*V. fordii*	99	100
51_Locus_1637_contig_2_3	0.714	5,004	7,686	3,631–4,791	Full	386			*V. fordii*	99	100
21_Locus_314_contig_13_20	0.155	1,531	242	1,400–24	Full	458	Omega-3 fatty acid desaturase	FAD3	*Jatropha curcas*	98	76
51_Locus_14571_contig_1_4	0.667	1,597	276	774–130	Incomplete	214			*V. fordii*	99	81
51_Locus_14571_contig_3_4	0.667	1,591	282	1,346–444	Incomplete	300	Omega-3 fatty acid desaturase-endoplasmic reticulum	FAD3 ER	*Ricinus communis*	89	78
31_Locus_4593_contig_4_4	0.769	2,076	493	1,789–431	Full	452	Omega-3 fatty acid desaturase precursor	FAD3 PRECURSOR	*V. fordii*	97	97
61_Locus_4845_contig_1_2	0.75	2,029	491	288–1,646	Full	452			*V. fordii*	97	97
41_Locus_15998_contig_1_1	1	1,740	136	132–1,502	Full	456			*R. communis*	99	86
31_Locus_1057_contig_3_10	0.172	997	605	802–164	Incomplete	212	Chloroplast omega-6 fatty acid desaturase	FAD6	*J. curcas*	94	83
41_Locus_1134_contig_5_9	0.667	2,085	1458	1,890–637	Full	417			*J. curcas*	93	88
21_Locus_244_contig_22_25	0.064	2,767	584	1,244–132	Full	370	Delta-9 fatty acid desaturase	FAD9	*V. fordii*	84	100

Contigs of the tung transcriptome were utilized for homology searches against non-redundant sequences from NCBI (http://www.ncbi.nlm.nih.gov/) and Swiss-Prot (http://www.expasy.ch/sprot/) by applying the BLASTX program (*e* value $<1e^{-6}$). The coverage and identity of the alignment, gene description, ORF length, size and structure, and species with homologous sequences are presented. The number of reads was calculated based on the alignment of the reads from RNA-Seq against the contigs

nt nucleotide, *aa* aminoacid

qPCR results, the expression of FADX increases more than 7,000-fold in mature seeds compared with seeds from stage 1 (Fig. 6), confirming its importance in tung seeds. As FADX is the only enzyme currently described as possessing the capability of introducing a double bond at the Δ12 position in the *trans* configuration, which is observed in α-eleostearic acid, the sequence and expression profile information of FADX will be useful in further studies aiming at the production of transgenic plants containing this industrially important fatty acid. In the present work, it was also possible to identify other fatty acid desaturases genes in *V. fordii* seeds. These include the full-length sequences of the FAD3 precursor (452 amino acids) and delta-9 fatty acid desaturase (FAD9; 370 amino acids), which were previously deposited as incomplete sequences in the GenBank database, as well as the homologous *J. curcas* chloroplastid omega-6 fatty acid desaturase (FAD6; 417 amino acids) and an incomplete sequence of a *Ricinus communis* FAD3 homolog from ER (300 amino acids) (Table 2). In addition to analyzing the expression profile of FAD2, FAD3, and FADX, we also investigated the expression profile of *V. fordii* FAD9 and FAD3 precursors for the first time (Fig. 6). The FAD3 precursor was constantly expressed during seed development; in contrast, the expression of FAD9 increased almost 15-fold in the mature seed compared with seeds from stage 1. The role of FAD9 as a key enzyme in the synthesis of polyunsaturated fatty acids in diatoms was recently described (Muto et al. 2013), and the expression of *Saccharomyces cerevisiae* FAD9 in soybean seeds resulted in the efficient conversion of 16:0 to 16:1Δ9 (Xue et al. 2013). These results suggest that tung FAD9 may also play an important role in the accumulation of tung oil or in the introduction of specific double bounds in fatty acids in tung seeds; therefore, *V. fordii* FAD9 must be further investigated in detail. A high content of unsaturated fatty acids in oil is undesirable for biodiesel production because unsaturated fatty acids affect the oxidative stability and ignition quality of biodiesel (Knothe 2005). Therefore, the desaturases and conjugase identified in the present study are potential candidates for RNAi constructs for the efficient modification of the fatty acid composition of tung seed oil to improve its fuel properties for use as biodiesel.

Activation and transport of fatty acids

To enable the transport of free fatty acids to the ER, activation to CoA esters by long-chain acyl-CoA synthetase (LACS) is required. Acyl-CoA binding proteins (ACBPs) then bind to the activated fatty acids to protect them from acyl-CoA hydrolases and to transport them to the ER. ACBP-bound fatty acids may also be converted to phosphatidyl choline (PC) at the plastid envelope by the action of lysophosphatidylcholine acyltransferase (LPCAT) (Fig. 5). In this study, we identified 29 unigenes for LCAS and 19 for ACBPs, which should be analyzed deeply.

Synthesis of triacylglycerol

The major component of plant seed oil is TAG, which acts as an energy reserve. Two metabolic pathways for the production of TAGs have been elucidated: an acyl-CoA-dependent pathway and an acyl-CoA-independent pathway, both occurring in the ER (Li-Beisson et al. 2013). In the acyl-CoA dependent pathway, commonly known as the Kennedy pathway, acyl-CoA is used as a substrate for the serial incorporation of three acyl groups into the glycerol backbone. This pathway is dependent on enzymes such as glycerol-3-phosphate acyltransferase (G3PAT), lysophosphatidic acid acyl transferase (LPAT), and phosphatidic acid phosphatase (PAP), resulting in the formation of diacylglycerol (DAG) (Fig. 5). Several homologs of G3PAT and LPAT have been identified (Xu et al. 2011; Gu et al. 2012). We found 19 unigenes coding for G3PAT, 11 for LPAT, and 36 for PAP. This large number of unigenes indicates that there may be multiple copies in *V. fordii* seeds that are differentially regulated. The expression of tung LPAT1 was evaluated during seed development (Fig. 6), showing a constant expression level during all stages and suggesting that other copies of LPAAT are of major importance for oil accumulation in tung seeds. The overexpression of two rapeseed LPAAT isozymes in Arabidopsis increased the seed lipid content (Maisonneuve et al. 2010), and the seed oil content was increased by 21 % in Arabidopsis and 3-7 % in canola by overexpressing GPAT (Jain et al. 2000) and DGAT (Taylor et al. 2009), respectively, suggesting that increasing the expression of glycerolipid acyltransferase in seeds leads to a greater flux of intermediates through the Kennedy pathway and enhanced TAG accumulation. These approaches could also be used to further increase the content of TAGs in tung seeds.

DAGs are converted to TAG by diacylglycerol acyltransferases (DAGTs). Three orthologs of DGATs have been found in tung tree. The DGAT1 and DGAT2 enzymes are unrelated, differing not only in their sequence and membrane topology but also in terms of their substrate discrimination. The expression of tung tree DGAT2 in yeast cells resulted in elevated accumulation of TAG compared with DGAT1 (Shockey et al. 2006). In addition, a third, soluble class of DGAT enzyme has been reported in tung (Cao et al. 2013). In the present study, we identified 30 unigenes for DGAT1 and 17 for DAGT2, and 3 unigenes were annotated as homologous of *R. communis* soluble DGAT. The RT-qPCR analysis of tung DGAT1 and DGAT2 showed little difference in the steady-state expression of DGAT1, with a small increase in the expression at the S4 stage, whereas DGAT2 was found to be highly expressed at the end of seed maturation (S4 and S5 stages, Fig. 6). These results confirmed those obtained by

Shockey et al. (2006), reporting that DGAT2 is the major DGAT mRNA in tung seeds and is correlated with oil accumulation. The same result was obtained for the developing seeds of Arabidopsis, soybean, Vernonia, Stokesia, Euphorbia, and castor bean (Kroon et al. 2006), indicating that DGAT2 preferentially incorporates unusual FAs (such as ricinoleate and α-eleostearic acid) into TAG, with DGAT1 being responsible for incorporating usual FAs (such as oleate and palmitate) into TAG within oleaginous seeds. Therefore, the availability of the full-length sequences of tung DGATs will be useful for increasing the oil content in *V. fordii* by genetic engineering.

The acyl-CoA-independent pathway for the synthesis of TAGs involves phospholipid/diacylglycerol acyltransferase (PDAT) to produce TAG. Choline phosphotransferase (CPT) can also contribute to TAG biosynthesis by catalyzing the reversible conversion of PC to DAG (Fig. 5). We identified nine unigenes annotated as PDAT and 6 as CPT (Table S6). PDAT genes are thought to play a role in the incorporation of unusual fatty acids into TAGs in castor bean (ricinoleate), *Vernonia galamensis*, *Euphorbia lagascae*, and *Stokesia laevis* (vernolic acid) because they show increased expression along seed development (Li et al. 2010). The expression profile of tung PDAT indicates that the expression level is constant in all seed developmental stages (Fig. 6). These results suggest that the acyl-CoA-independent pathway may not be the most important pathway for the synthesis of TAGs in tung seeds.

Oil storage proteins in tung seeds

Once synthesized, pools of TAG can be stored in the mature seed in the form of oil bodies or lipid droplets, which are surrounded by a layer of phospholipids with a number of proteins. The most abundant of these proteins are known as oleosins (OLE), but caleosins and steroleosins also exist (Fig. 5). The OLE are thought to stabilize the oil body during desiccation of the seed and determine the size of the oil bodies and therefore may regulate oil storage (Voelker and Kinney 2001). Considering that the suppression of OLE had a small but statistically significant effect on fatty acid preferences for TAG (Siloto et al. 2006), it is supposed that the introduction of a foreign OLE may be an alternative way to select a particular fatty acid and increase TAG accumulation through modulation of the oil body size. Sequences of OLE in *V. fordii* were previously identified by Chen et al. (2010). These authors identified 118 ESTs, approximately 4.3 % of the total sequenced ESTs, corresponding to OLE genes. In our study, we found 15 ESTs coding for OLE, 16 for calosins, and 10 for steroleosins (Table S6), suggesting that these proteins play a role in oil bodies. The expression profile of OLE1 was evaluated by RT-qPCR and showed that OLE1 increases its expression during seed development, positively correlating with oil accumulation in the seed (Fig. 6). The expression of OLE genes was strongly upregulated and tightly correlated with TAG biosynthesis in the developing seeds of rapeseed (Jolivet et al. 2011), castor (Chen et al. 2007), and Jatropha (Xu et al. 2011), suggesting that the two processes of OLE and TAG accumulation are closely linked.

Metabolic pathways related to oil breakdown in tung seeds

TAGs must be hydrolyzed during germination to be used by the embryo as an energy source. This process is mediated by TAG lipases (TL), resulting in free fatty acids and the

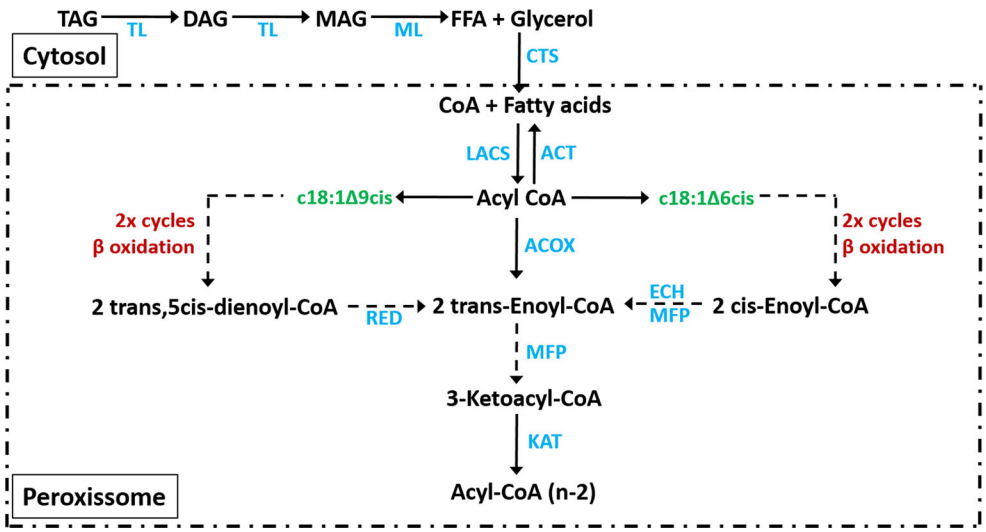

Fig. 7 Oil breakdown pathway in tung seeds. The enzymes for which genes were identified in the tung transcriptome are shown in *light blue*. *TAG* triacylglycerol, *TL* triacylglycerol lipase, *DAG* diacylglycerol, *MAG* monoacylglycerol, *ML* monoacylglyceride lipase, *FFA* free fatty acid, *CTS* ABC transporter involved in the import of b-oxidation substrates to the peroxisome, *LACS* long-chain acyl-CoA synthetase, *ACT* acyl-CoA thioesterase, *ACOX* acyl-CoA oxidase, *RED* 2,4-dienoyl-CoA reductase, *ECH* 2E-enoyl-CoA hydratase, *MFP* multifunctional protein, *KAT* 3-ketothiolase

intermediate products diacylglycerol or monoacylglycerol (Voelker and Kinney 2001; Lu et al. 2007). As expected, we identified a large number of TL and monoglyceride lipases (ML) in our transcriptome (Table S7). Based on this analysis, a pathway for oil metabolism in *V. fordii* is proposed in Fig. 7.

According to Fig. 7, the fatty acids released by TAG lipolysis are transported across the peroxisome membrane by an ABC transporter protein (CTS). Then, *o*-succinylbenzoate-CoA ligase (LACS) is responsible for the esterification of the fatty acids into acyl-CoA moieties that are used in two cycles of β-oxidation, in which acetyl-CoA is sequentially cleaved from acyl-CoA. This process requires the enzymes acyl-CoA oxidase (ACOX), multifunctional protein (MFP), and 3-ketoacyl-CoA thiolase (KAT) to catalyze the oxidation, hydration and dehydrogenation, and thiolytic cleavage, respectively, of acyl-CoA. Furthermore, many fatty acids have unsaturated bonds in the *cis*-configuration that result in metabolic blocks for the β-oxidation pathway. Therefore, enoyl-CoA hydratase2 (ECH) produces trans-enoyl-CoA from C18:1 Δ9 *cis*, which re-enters the normal set of core reactions of β-oxidation (Fig. 7). We found unigenes coding for all enzymes in the fatty acid degradation pathway. These unigenes are shown in Table S7, including information regarding length, abundance, ORF structure, and the species with the best-hit homologous sequence.

Conclusions

In the present work, we provide a collection of unigenes derived from mature tung seeds. We performed assembly and annotation of the transcriptome, resulting in the identification of the sequence of enzymes and proteins from all steps of oil biosynthesis and breakdown in tung seeds. The expression profile of a subset of those genes was analyzed during seed development, indicating that BCCP, FAD2, FAD3, FADX, FAD9, DGAT2, and OleoI are highly expressed at the end of seed development, which correlates with oil accumulation. The precise role of these genes and their paralogs with potential neo- or subfunctionalization will require further studies using functional genomic approaches. These genes are promising targets for metabolic engineering efforts. Further studies regarding the overexpression of enzymes related to oil synthesis, such as BCCP and OleoI, associated with the downregulation of enzymes related to oil breakdown, and to the production of unsaturated fatty acids, such as FAD2, FAD3, FAD9, and FADX, identified in this study, may result in the production of a high content of oil in tung seeds with high quality for use as biodiesel. Furthermore, the introduction of tung desaturases and conjugases into more productive crops may enable them to accumulate unsaturated and unusual fatty acids largely for use in industrial applications.

Acknowledgments This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grant number 559636/2009-1, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo a Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Financiadora de Projetos (FINEP), Embrapa, Petrobrás, and Ministério de Ciência e Tecnologia (MCT).

Data Archiving Statement The RNA-Seq data and the transcriptome of tung tree seeds are currently been submitted to GenBank as Gene Expression Omnibus (GEO) data libraries. The accession numbers will be supplied once available.

References

Allen CAW, Watts KC, Ackman RG, Pegg MJ (1999) Predicting the viscosity of biodiesel fuels from their fatty acid ester composition. Fuel 78:1319–1326

Andre C, Haslam RP, Shanklin J (2012) Feedback regulation of plastidic acetyl-CoA carboxylase by 18:1-acyl carrier protein in *Brassica napus*. PNAS 109:10107–10112

Banik M, Duguid S, Cloutier S (2011) Transcript profiling and gene characterization of three fatty acid desaturase genes in high, moderate, and low linolenic acid genotypes of flax (*Linum usitatissimum L.*) and their role in linolenic acid accumulation. Genome 54:471–483

Brown K, Keeler W (2005) The history of tung oil. Wildland Weeds 9:4–24

Cahoon EB, Dietrich CR, Meyer K, Damude HG, Dyer JM, Kinney AJ (2006) Conjugated fatty acids accumulate to high levels in phospholipids of metabolically engineered soybean and Arabidopsis seeds. Phytochem 67:1166–1176

Cao H, Shockey JM, Klasson KT, Chapital DC, Mason CCB, Scheffler BE (2013) Developmental regulation of diacylglycerol acyltransferase family gene expression in tung tree tissues. Plos ONE 8: e76946

Chandran D, Sankararamasubramanian HM, Kumar MA, Parida A (2014) Differential expression analysis of transcripts related to oil metabolism in maturing seeds of *Jatropha curcas L*. Physiol Mol Biol Plants 20:181–190

Chen GQ, Turner C, He X, Nguyen T, McKeon TA, Laudencia-Chingcuanco D (2007) Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in castor bean (*Ricinus communis L.*). Lipids 42:263–274

Chen Y, Zhou G, Wang Y, Xu L (2010) F-BOX and oleosin: additional target genes for future metabolic engineering in tung trees? Ind Crop Prod 32:684–686

Conesa S, Götz (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 1:12

Crawford JE, Guelbeogo WM, Sanou A, Traoré A, Vernick KD, Sagnon N, Lazzaro BP (2010) De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-Sseq technology. PLoS One 5:e14202

Dyer JM, Mullen RT (2005) Development and potential of genetically engineered oilseeds. Seed Sci Res 15:255–267

Dyer JM, Chapital DC, Kuan JCW, Mullen RT, Turner C, McKeon TA, Pepperman AB (2002) Molecular analysis of a bifunctional fatty acid conjugase/desaturase from tung implications for the evolution of plant fatty acid diversity. Plant Physiol 130:2027–2038

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461

Garg R, Patel RA, Khilesh A, Tyagi K, Jain MU (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res 18:53–63

Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhart PJ (2012) Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics 13:92

Gu K, Yi C, Tian D, Sangha JS, Hong Y, Yin Z (2012) Expression of fatty acid and lipid biosynthetic genes in developing endosperm of Jatropha curcas. Biotechnol Biofuels 5:47

Hu Y, Wu G, Cao Y, Wu Y, Xiao L, Li X, Lu C (2009) Breeding response of transcript profiling in developing seeds of Brassica napus. BMC Mol Biol 10:1–17

Jain RK, Coffey M, Lai K, Kumar A, MacKenzie SL (2000) Enhancement of seed oil content by expression of glycerol-3-phosphate acyltransferase genes. Biochem Soc Trans 28: 958–961

Jaworski J, Cahoon EB (2003) Industrial oils from transgenic plants. Curr Opin Plant Biol 6:178–184

Jiang H, Wu P, Zhang S, Song C, Chen Y, Li M, Jia Y, Fang X, Chen F, Wu G (2012) Global Analysis of gene expression profiles in developing physic nut (*Jatropha curcas L.*) seeds. Plos One 7:e36522

Jolivet P, Boulard C, Bellamy A, Valot B, d'Andréa S, Zivy M, Nesi N, Chardot T (2011) Oil body proteins sequentially accumulate throughout seed development in Brassica napus. J Plant Physiol 168:2015-2020

Ke J, Wen TN, Nikolau BJ, Wurtele ES (2000) Coordinate regulation of the nuclear and plastidic genes coding for the subunits of the heteromeric acetyl-coenzyme a carboxylase. Plant Physiol 122: 1057–1071

Knothe G (2005) Dependence of biodiesel fuel properties on the structure of fatty acid alkyl esters. Fuel Process Technol 86:1059–1070

Kroon JTM, Wei WX, Simon WJ, Slabas AR (2006) Identification and functional expression of a type 2 acyl-CoA: diacylglycerol acyltransferase (DGAT2) in developing castor bean seeds which has high homology to the major triglyceride biosynthetic enzyme of fungi and animals. Phytochem 67:2541–2549

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li R, Yu K, Hildebrand DF (2010) DGAT1, DGAT2 and PDAT expression in seeds and other tissues of epoxy and hydroxy fatty acid accumulating plants. Lipids 45:145–157

Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, Bates PD, Baud S, Bird D, DeBono A, Durrett TP et al (2013) Acyl-lipid metabolism. Arabidopsis Book 11:e0161

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCt Method. Methods 25:402–408

Lu CF, Wallis JG, Browse J (2007) An analysis of expressed sequence tags of developing castor endosperm using a full-length cDNA library. BMC Plant Biol 7:42

Maisonneuve S, Bessoule JJ, Lessire R, Delseny M, Roscoe TJ (2010) Expression of rapeseed microsomal lysophosphatidic acid acyltransferase isozymes enhances seed oil content in Arabidopsis. Plant Physiol 152:670–684

Meier S, Tzfadia O, Vallabhaneni R, Gehring C, Wurtzel ET et al (2011) A transcriptional analysis of carotenoid, chlorophyll and plastidial isoprenoid biosynthesis genes during development and osmotic stress responses in *Arabidopsis thaliana*. BMC Syst Biol 5:77

Muto M, Kubota C, Tanaka M, Satoh A, Matsumoto M, Yoshino T, Tanaka T (2013) Identification and functional analysis of delta-9 desaturase, a key enzyme in PUFA synthesis, isolated from the oleaginous diatom fistulifera. PlosOne 8:e73507

Nakkaew A, Chotigeat W, Eksomtramage T, Phongdara A (2008) Cloning and expression of a plastid-encoded subunit, beta-carboxyltransferase gene (accD) and a nuclear-encoded subunit, biotin carboxylase of acetyl-CoAcarboxylase from oil palm (*Elaeis guineensis* Jacq.). Plant Sci 175:497–504

Natarajan P, Parani M (2011) De novo assembly and transcriptome analysis of five major tissues of *Jatropha curcas* L. using GS FLX titanium platform of 454 pyrosequencing. BMC Genomics 12:191

Nguyen T, Shanklin J (2009) Altering arabidopsis oilseed composition by a combined antisense-hairpin RNAi gene suppression approach. J Am Oil Chem Soc 86:41–49

Park JY, Kim DK, Wang ZM, Lu P, Park SC, Lee JS (2008) Production and characterization of biodiesel from tung oil. Appl Biochem Biotechnol 148:109–117

Rajwade, AV, Kadoo NY, Borikar SP, Harsulkar AM, Ghorpade PB, Gupta VS (2014) Differential transcriptional activity of SAD, FAD2 and FAD3 desaturase genes in developing seeds of linseed contributes to varietal variation in a-linolenic acid content. Phytochem 98:41–53

Rupilinus W, Ahmad S (2007) Palm oil and palm kernel oil as raw materials for basic oleochemicals and biodiesel. Eur J Lipid Sci Technol 109:433–439

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864

Schmucki R, Berrera M, Küng E, Lee S, Thasler WE, Grüner S, Ebeling M, Cert U (2013) High throughput transcriptome analysis of lipid metabolism in Syrian hamster liver in absence of an annotated genome. Genomics 14:237

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNAseq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092

Shang Q, Jiang W, Lu H, Liang B (2010) Properties of tung oil biodiesel and its blends with diesel. Bioresour Technol 101:826–828

Shockey JM, Gidda SK, Chapital DC, Kuan JC, Dhanoa PK et al (2006) Tung tree DGAT1 and DGAT2 have nonredundant functions in triacylglycerol biosynthesis and are localized to different subdomains of the endoplasmic reticulum. Plant Cell 18:2294–2313

Siloto RMP, Findlay K, Lopez-Villalobos A, Yeung EC, Nykiforuk CL, Moloney MM (2006) The accumulation of oleosins determines the size of seed oilbodies in Arabidopsis. Plant Cell 18:1961–1974

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res 20:1432–1440

Taylor DC, Zhang Y, Kumar A, Francis T, Giblin EM, Barton DL, Ferrie JR, Laroche A, Shah S, Zhu W, Snyder CL, Hall L, Rakow G, Harwood JL, Weselake RJ (2009) Molecular modification of triacylglycerol accumulation by over-expression of DGAT1 to produce canola with increased seed oil content under field conditions. Botany 87:533–543

Vanhercke T, Craig CW, Stymne S, Singh SP, Green AG (2013) Metabolic engineering of plant oils and waxes for use as industrial feedstocks. Plant Biotechnol J 11:197–210

Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Santo SD, Minio A, Buson G, Tononi P, Zago ED, Zamperin G, Bellin D, Pezzotti M, Delledonne M (2013) De novo transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. BMC Genomics 14:41

Voelker T, Kinney AJ (2001) Variations in the biosynthesis of seed-storage lipids. Annu Rev Plant Physiol Plant Mol Biol 52:335–361

Wada M, Yasuno R, Jordan SW, Cronan JE, Wada H (2001) Lipoic acid metabolism in Arabidopsis thaliana: Cloning and characterization of a cDNA encoding lipoyltransferase. Plant Cell Physiol 42:650–656

Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11:400

Xu R, Wang R, Liu A (2011) Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in developing seeds of Jatropha (*Jatropha curcas* L.). Biomass Bioenergy 35:1683–1692

Xue J, Mao X, Yang Z, Wu Y, Jia X, Zhang L, Yue A, Wang J, Li R (2013) Expression of yeast acyl-CoA-D9 desaturase leads to accumulation of unusual monounsaturated fatty acids in soybean seeds. Biotechnol Lett 35:951–959

Yang W, Qi Y, Bi K, Fu J (2012) Toward understanding the genetic basis of adaptation to high-elevation life in poikilothermic species: a comparative transcriptomic analysis of two ranid frogs, *Rana chensinensis* and *R. kukunoris*. BMC Genomics 13:588

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34:W293–W297

OPEN ACCESS Freely available online

PLOS ONE

Identifying MicroRNAs and Transcript Targets in *Jatropha* Seeds

Vanessa Galli[1,2], Frank Guzman[3], Luiz F. V. de Oliveira[1], Guilherme Loss-Morais[1], Ana P. Körbes[3], Sérgio D. A. Silva[2], Márcia M. A. N. Margis-Pinheiro[3], Rogério Margis[1,3]*

1 Center of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Populations, Federal University of Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil, **2** Brazilian Agricultural Research – EMBRAPA, Pelotas, RS, Brazil, **3** PPGGBM at Federal University of Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brazil

Abstract

MicroRNAs, or miRNAs, are endogenously encoded small RNAs that play a key role in diverse plant biological processes. *Jatropha curcas* L. has received significant attention as a potential oilseed crop for the production of renewable oil. Here, a sRNA library of mature seeds and three mRNA libraries from three different seed development stages were generated by deep sequencing to identify and characterize the miRNAs and pre-miRNAs of *J. curcas*. Computational analysis was used for the identification of 180 conserved miRNAs and 41 precursors (pre-miRNAs) as well as 16 novel pre-miRNAs. The predicted miRNA target genes are involved in a broad range of physiological functions, including cellular structure, nuclear function, translation, transport, hormone synthesis, defense, and lipid metabolism. Some pre-miRNA and miRNA targets vary in abundance between the three stages of seed development. A search for sequences that produce siRNA was performed, and the results indicated that *J. curcas* siRNAs play a role in nuclear functions, transport, catalytic processes and disease resistance. This study presents the first large scale identification of *J. curcas* miRNAs and their targets in mature seeds based on deep sequencing, and it contributes to a functional understanding of these miRNAs.

Citation: Galli V, Guzman F, de Oliveira LFV, Loss-Morais G, Körbes AP, et al. (2014) Identifying MicroRNAs and Transcript Targets in *Jatropha* Seeds. PLoS ONE 9(2): e83727. doi:10.1371/journal.pone.0083727

Editor: Steven George Rozen, Duke-NUS, Singapore

Received May 12, 2013; **Accepted** November 6, 2013; **Published** February 13, 2014

Funding: This research was supported by Conselho Nacional de Desenvolvimento Científico e Tencológico - CNPq grants numbers (559636/2009-1 and 307868/2011-7), Ministerio de Ciência e Tecnologia - MCT and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior- CAPES. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: I have read the journal's policy and have the following conflicts: We want to inform that the co-author Rogerio Margis is a PLOS ONE Editorial Board member. This fact does not alter our adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: rogerio.margis@ufrgs.br

Introduction

Increased energy consumption is a key concern of contemporary society as a result of rapidly diminishing fossil fuel deposits and increasing levels of carbon dioxide released into the atmosphere. Thus, environmentally friendly sources of fuels, such as bioethanol and biodiesel, are promising alternatives for fossil fuels. In this context, great interest has been generated regarding the potential of *Jatropha curcas* L. for biodiesel production. This species belongs to the Euphorbiaceae family and is found in almost all tropical areas; it occurs on a large scale in tropical and temperate regions [1,2]. *J. curcas* has potential for biodiesel production because it is perennial, drought-resistant and has a high oil content (~40%). Additionally, this crop can be grown in degraded soils (non-agricultural lands), which controls erosion without competing for food production habitats [3,4].

In spite of having the potential for high fuel production, improving the quality of *J. curcas* seed oil remains challenging. The desired traits include increased oil content, decreased unsaturated fatty acid content (to increase the oxidative stability), decreased free fatty acid content (to prevent soap formation and increase biodiesel productivity) and decreased 18 carbon fatty acid content (to reduce viscosity) [5]. Furthermore, reducing seed toxicity and increasing pest tolerance are also desirable [1,5].

MicroRNAs (miRNAs) are small non-coding RNAs that act as post-transcriptional regulators of gene expression [6]. They are typically transcribed by RNA Polymerase II as long polyadenylated transcripts, with an imperfect stem-loop structure known as pri-miRNA, which is recognized and processed by DICER-Like1 (DCL1) into an miRNA precursor (pre-miRNA). This sequence is further processed to generate mature miRNAs, which are composed of single-stranded RNA molecules of approximately 21 nucleotides (nt) in length [7–9]. To regulate protein-coding genes, the mature miRNA binds to sites in the 3′ and 5′ untranslated regions (UTR) or the coding sequence (CDS), leading to mRNA degradation or translational inhibition, depending on the degree of complementarity between the miRNA and its target transcript [8,10]. This regulation plays critical roles in plant development and growth as well as a range of physiological processes, including abiotic and biotic stress responses [11,12], and in lipid metabolism [13,14].

At present, there have only been two recent studies regarding the identification of miRNAs from *J. curcas*. One study used a small RNA cloning methodology and did not provide the precursor sequence from Jatropha [15]. The other study used known plant miRNAs from Viridiplantae to search for *J. curcas* miRNAs in publicly available EST and GSS databases; therefore, only conserved miRNAs were identified by this approach [16]. In the present study, conserved and novel miRNAs from *J. curcas* were identified through the deep sequencing of small RNAs (sRNA) from mature seeds. The targets of these miRNAs were

predicted *in silico*, and the results indicate that miRNAs are involved in a wide range of physiological processes in seeds, including growth, hormone signaling, stress resistance and lipid metabolism, among others. In addition, polyadenylated transcript sequencing (mRNA-seq) libraries from three seed development stages (immature, intermediate and mature) were used to identify and characterize the abundance of pre-miRNAs and miRNA targets, providing important information related to regulatory timing.

Material and Methods

J. curcas Seed Collection and RNA Isolation

For the RNA isolation, fruits from *J. curcas* plants grown in an open environment at Embrapa Clima Temperado (Pelotas, RS, Brazil) were collected at 10–20 (immature seeds), 20–40 (intermediate seeds) and 40–60 (mature seeds) days after flower opening (DAF). The seeds were dissected from their fruits and immediately frozen in liquid nitrogen and then stored at −80°C. Total RNA was isolated from a pool of seeds from each stage with Trizol (Invitrogen, CA, USA), according to the manufacturer's protocol. RNA quality was evaluated by electrophoresis on a 1% agarose gel, and the RNA concentration was determined by Nanodrop (Nanodrop Technologies, Wilmington, DE, USA).

Small RNA and mRNA-seq Library Construction using Deep Sequencing

Total RNA (>10 μg) was isolated from immature, intermediate and mature seeds and sent to Fasteris SA (Plan-les-Ouates, Switzerland) for processing and sequencing with an Illumina HiSeq2000. To generate the mRNA-seq libraries, polyadenylated transcript sequencing was performed as follows: Poly(A) mRNAs were purified, cDNA was synthesized using a Poly(T) primer shotgun to generate 500 nt inserts, 3p and 5p adapters were bound, and a cDNA colony template library was generated by PCR amplification, and 100 single-end bases were sequenced by Illumina sequencing. Illumina sequencing output data were sequence tags of 100 bases. Three mRNA-seq libraries were constructed from immature seeds (L1), from intermediate seeds (L2) and from mature seeds (L3).

Only RNA from mature seeds was used to produce the small RNA library. In brief, RNA bands corresponding to a size range of 20–30 nt were separated and purified from the acrylamide gel and subsequently bound to 3p and 5p adapters in two separate subsequent steps, each followed by acrylamide gel purification. Then, the cDNAs were synthesized and a cDNA colony template library was generated by PCR amplification for Illumina sequencing.

Sequence Data Analysis

Figure S1 summarizes the overall data analyses performed with the sRNA library and mRNA-seq libraries. First, all low quality reads (FASTq value <13) were removed, and 5p and 3p adapter sequences were trimmed using the Genome Analyzer Pipeline (Fasteris). The remaining low quality reads with 'n' were removed using the PrinSeq script [17]. Sequences shorter than 18 nt and longer than 25 nt were excluded from further analysis. sRNAs derived from Viridiplantae rRNAs, tRNAs, snRNAs and snoRNAs (from the tRNAdb [18]; SILVA rRNA [19] and NONCODE v3.0 [20] databases); cpRNA from *J. curcas*; and mtRNA from *Ricinus communis* [from the NCBI GenBank database (http://ftp.ncbi.nlm.nih.gov)] were identified by mapping with Bowtie v 0.12.7 [21]. rRNAs, tRNAs, snRNAs and snoRNAs were excluded from the sRNA library that was used in further miRNA predictions and analyses.

The mRNA-seq libraries were first filtered for low quality "n" reads, and *J. curcas* coding sequence reads were obtained from the *Jatropha* genome database (http://www.kazusa.or.jp/jatropha). L1, L2 and L3 were pooled to produce mRNA contigs using the CLC Genome Workbench version 4.0.2 (CLCbio, Aarhus, Denmark) algorithm for *de novo* sequence assembly, with the default parameters (similarity = 0.8, length fraction = 0.5, insertion/deletion cost = 3, mismatch cost = 3), originating from the L1–L2–L3 library.

Sequence data from this article can be found in the GenBank data libraries under accession number(s) GSM1226039 (L1), GSM1226040 (L2), GSM1226038 (L3) and GSM1226041 (sRNA dataset).

Prediction of Conserved and Novel miRNAs

To identify phylogenetically conserved miRNAs, reads from the sRNA library derived from mature seeds were mapped to a set of all mature Viridiplantae unique miRNAs obtained from the miRBase database (Release 19, August 2012) using Bowtie v 0.12.7 [21]. Only perfectly matched sequences were considered to be known miRNAs. To search for novel miRNAs, reads from the sRNA library derived from mature seeds were matched against contigs assembled from the L1–L2–L3 library using SOAP2 [22]. The SOAP2 output was filtered with an *in house* filter tool to separate pre-miRNA candidate sequences by using an anchoring pattern of one or two blocks of aligned sRNAs with a perfect match. MFOLD (http://mfold.bioinfo.rpi.edu/cgi-bin/rnaform1.cgi) was employed to predict hairpin structures with default parameters. Sequences were considered to be pre-miRNA if the RNA sequence could form an appropriate stem-loop structure with a mature miRNA sitting in one arm of the hairpin structure; the secondary structure had a high negative minimum folding free energy (MFE, 17–110 kcal/mol), using RNAstructure 5.3 [23], and a high negative minimum folding free energy index (MFEI, higher than 0.5). All putative pre-miRNAs were verified by a BLASTn algorithm from NCBI databases and the miRBase database (Release 19, August 2012). The frequency of identified miRNAs was obtained by aligning the conserved and novel precursors identified in this study and the sRNA library using Bowtie v 0.12.7, with the default parameters. The SAM files from Bowtie were then processed using *in house* Python scripts to count the frequencies of each read and map them into the three libraries. The most frequent miRNA for each precursor was designated as miRNA, while the others were designated as isomiRNAs.

miRNA Targets Prediction

mRNA contigs from the L1–L2–L3 library were clustered by using Gene Indices clustering tools (http://compbio.dfci.harvard.edu/tgi/software/) [24] to reduce sequence redundancy. The clustering output was passed on to a CAP3 assembler [25] for multiple alignment and consensus building. Contigs that could not reach the threshold set and fell into a random assembly and remained as a list of singletons.

The predicted target genes of conserved and novel miRNAs was performed with a psRNAtarget [26] by aligning mature sequences against assembled *J. curcas* unigenes from the L1–L2–L3 library. Default parameters were used, and a maximum expectation of 4.0 was applied for the search with the most abundant miRNA. A maximum expectation of 5.0 was used to search the isomiRNA target genes, which were related to lipid metabolism pathways. An annotation of predicted targets was performed by using BLASTX from Blast2GO v2.3.5 software [27] based on their sequence

similarity with previously identified and annotated genes from the NR and Swiss-Prot/Uniprot protein databases. The annotation was improved by analyzing conserved domains/families using the InterProScan tool, and Gene Ontology (GO) terms for the cellular component, molecular function and biological processes were determined by using the GOslim tool in the blast2GO software. Transcript orientations were obtained from the BLAST outputs.

In silico Expression Analysis of Pre-miRNAs and miRNA Predicted Targets

To calculate the frequency of pre-miRNAs, mRNA reads from each individual library (L1, L2 and L3) were aligned in Bowtie v 0.12.7 by using the default parameters and allowing zero mismatches. For reference, all identified pre-miRNAs in this study were used. The SAM files from Bowtie were then processed, as discussed above. The scaling normalization method proposed by Robinson and Oshlack [28] was used for data normalization. Both R packages, EdgeR [29] and A–C test [30], were independently used to assess whether the pre-miRNA was differentially represented. In brief, EdgeR uses a negative binomial model to estimate the over dispersion from the pre-miRNA count. The dispersion parameter of each pre-miRNA was estimated by tagwise dispersion. The differential expression is then assessed for each pre-miRNA by using an adapted exact test for over dispersed data. The A-C test computes the probability that two independent counts of the same pre-miRNA came from similar samples. Pre-miRNAs were considered to be differentially represented if they had a p-value ≤ 0.001 in both statistical tests. The same method was adopted to evaluate the expression profile of predicted miRNAs targets, allowing two mismatches (one in the seed and another in the rest of the sequence).

siRNA Prediction

siRNAs were identified by aligning *J. curcas* 24-nt sRNAs against the contigs from the L1–L2–L3 library. Putative contigs with a typical sRNA distribution pattern along the matching sequences [31] were further subjected to annotation using Blast2GO software, as described above.

Results and Discussion

Deep Sequencing of sRNAs and mRNA Libraries

To identify the conserved and novel miRNAs in *J. curcas* seeds, an sRNA library from mature seeds was constructed and sequenced by Illumina technology, resulting in a total of 16,771,931 reads. After removing the 3p and 5p adapter sequences and filtering out low quality "n" sequences, sRNAs within a 1–44 nt range were obtained, in which the majority were 18–26 nt in length (Table 1). Sequences shorter than 18 nt and longer than 25 nt were removed, resulting in 13,953,403 reads (Table 2), from which 5,400,278 reads were unique tags (38.7%). Approximately 80% (4,328,139) of the unique tags corresponded to singletons (Table S1).

Non-coding RNAs were also removed from the sRNA library for further analysis (Table 2). The sequence analysis showed that rRNA had the highest read frequency of all filtered sRNA classes, with 6.94% of the total reads. The majority of these rRNA sequences were found in the dataset with 21 nt-long sequences. Interestingly, 35.95% of the 18 nt sequences represented rRNAs. cpRNA sequences were the second most abundant filtered sequences after those from rRNA, corresponding to 1.9% of the total reads. tRNAs, mtRNAs, snRNAs and snoRNA were less frequent in the sRNA library. Taken together, 9.71% of the sRNA library was filtered with these RNA types, leaving 12,597,985 reads.

Table 1. Raw data from sequencing *Jatropha curcas* sRNAs.

Read length (nt)	Number of reads	Percentage (%) of reads
0	12,163	0.073
1–17	1,347,333	8.330
18–26	14,175,504	84.190
27–44	516,355	3.790
Remaining	720,576	4.960

doi:10.1371/journal.pone.0083727.t001

The length distribution of redundant and non-redundant sRNAs reads indicated that the most abundant and diverse sequences are within 21 (20.89%) and 24 nt (44.55%), a typical size range for Dicer-like (DCL)-derived products [9]. This distribution pattern for the small RNA size is similar to that of seeds from other species, such as Arabidopsis [32], peanut [13], barley [33], soybean [34] and canola [14], which implies that *J. curcas* possesses similar small RNA biogenesis processing components to other plant species. The same length distribution pattern was observed before and after filtering the sRNA library (Figure 1), indicating that the small RNA library was of high quality.

In this study, a specific mRNA transcriptome of *J. curcas* seeds was produced and used as a sequence reference for further analyses. Three mRNA seqs were obtained from seeds as follows: the L1 (immature seed) library yielded 43,328,830 reads, the L2 (intermediate mature) library yielded 35,062,185 reads, and the L3 (mature seed) library yielded 16,653,188 reads. All three libraries were pooled for *de novo* assembly (L1–L2–L3). The resulting 61,863 contigs had an average length of 755 bp, and the size ranged between 100 bp and 15,706 bp, with 27,520 contigs constituting more than 500 bp in length.

Identification of Conserved miRNAs and Pre-miRNAs

Several studies have reported miRNA conservation across different plant taxa [35,36]. To identify homologous miRNAs in the *J. curcas* sRNA library, a set of 2,585 unique mature plant miRNA sequences were extracted from the miRBase database and used for alignment against the sRNA library. Only sequences with exactly the same size and nucleotide composition were considered. A strict criterion of sharply defined distribution patterns for one or two block-like anchored sRNAs and at least 10 reads of a single miRNA sequence were used to predict novel miRNAs (see methods). The read depth distribution along putative pre-miRNAs was previously shown to be a reliable guide for differentiating possible miRNAs from contaminant sequences, such as the degradation products of mRNAs or transcripts that are simultaneously expressed in both sense and antisense orientations [14,33,37]. In total, 1,021,895 reads perfectly matched 177 conserved miRNAs belonging to 41 families, with an average of approximately 4 miRNA members per family (Table S2). Overall, the Jcu_MIR167 family was the most abundant conserved miRNA family present in *J. curcas* seeds, accounting for 842,066 reads, and the largest families were Jcu_MIR156 and Jcu_MIR166, with 23 and 21 members, respectively. Of the remaining miRNA families, 19 contained between 2 to 8 members, and 16 were represented by a single member (Figure 2 and Table S2). Furthermore, the results indicate that different members of the same miRNA family have clearly different expression levels (Table S2). For example, Jcu_MIR166 presents members ranging from 1 to 35,439 reads.

Table 2. Data from reads of the sRNA database of mature *Jatropha curcas* seeds filtered with non-coding RNAs.

Size	Total reads*	%	rRNA		tRNA		snRNA		snoRNA		cpRNA		mtRNA		all filters	
			total reads	%	total reads	%	total reads	%	total reads	%	total reads	%	total reads	%	total reads	%
18	369,493	2.65	132,827	35.95	9,763	2.64	283	0.08	167	0.05	6,069	1.64	2,269	0.61	151,378	40.97
19	445,999	3.20	131,575	29.50	22,091	4.95	241	0.05	130	0.03	10,314	2.31	2,043	0.46	166,394	37.31
20	503,781	3.61	107,113	21.26	14,525	2.88	240	0.05	102	0.02	27,221	5.40	2,534	0.50	151,735	30.12
21	2,869,361	20.56	164,287	5.73	5,755	0.20	959	0.03	78	0.00	51,618	1.80	14,261	0.50	236,958	8.26
22	2,271,978	16.28	117,565	5.17	7,860	0.35	374	0.02	61	0.00	106,135	4.67	3,714	0.16	235,709	10.37
23	1,335,058	9.57	125,303	9.39	7,860	0.59	257	0.02	33	0.00	23,003	1.72	2,279	0.02	158,735	11.89
24	5,760,030	41.28	102,477	1.78	9,347	0.16	356	0.01	27	0.00	31,370	0.54	3,447	0.06	147,024	2.55
25	397,703	2.85	87,646	22.04	6,541	1.64	178	0.04	17	0.00	8,911	2.24	4,192	1.05	107,485	27.03
Total	13,953,403	100.00	968,793	6.94	83,742	0.60	2,888	0.02	615	0.00	264,641	1.90	34,739	0.25	1,355,418	9.71

*Total reads before filtering with non-coding and organellar small RNAs. The small RNAs were clustered according to their origin as follows: ribosome (rRNA), transporter (tRNA), small nuclear (snRNA), small nucleolar (snoRNA), mitochondrial (mtRNA) and chloroplastic (cpRNA).

doi:10.1371/journal.pone.0083727.t002

Interestingly, the conserved miRNAs represented the most abundant *J. curcas* miRNAs and were distributed throughout seven families (MIR156, MIR157, MIR159, MIR166, MIR167, MIR168 and MIR396). These abundant miRNA families are largely found in Viridiplantae, indicating a fundamental role in plant life maintenance (Table S2).

In a study performed by Wang et al. [15], six conserved miRNAs were identified by the cloning approach from RNA libraries derived from theleaves and developing seeds of *J. curcas*. Two (Jcu_MIR166_3p and Jcu_MIR167_5p) of them were also identified in the present study, and three conserved miRNAs were identified only in the library from leaves. Moreover, the miRNA JcumiR004, which was identified as a novel plant miRNA, according to Wang et al. [15], was annotated in the present study as Jcu_MIR171_5p because the mature sequence showed a perfect match with the MIR171 from several species (Table S2, sequence UGAUUGAGCCGUGCCAAUAUC). Therefore, the discrepancies in identifying conserved miRNAs from the present study and the one performed by Wang et al. [15] correspond mostly to a difference in selected tissues and methods. During the cloning approach, there was a chance miRNAs with low expression level would not be detected. A more recent work by Vishwakarma and Jadeja [16] also focused on the identification of conserved miRNAs from *J. curcas* after transcript and partial genome sequence analysis. These authors were able to identify 24 predicted miRNAs belonging to five miRNA families (Jcu_-MIR166, Jcu_MIR167, Jcu_MIR1096, Jcu_MIR5368 and Jcu_-MIR5021). A lower number of miRNA families were identified by these authors relative to the present study, most likely because they used known plant miRNAs from Viridiplantae to search the conserved *J. curcas* miRNAs homologs in publicly available (and relatively small) EST and GSS databases compared to the database from the RNAseq generated in the present study.

To identify putative conserved pre-miRNA sequences, the sRNA library was matched against a set of *de novo* assembled contigs from three developmental stages of *J. curcas* seeds (L1–L2–L3 library). The candidate pre-miRNAs were predicted by exploring the secondary structure, the minimum folding free energy (MFE) and the minimum folding free energy index (MFEI). Candidate mRNA sequences with a stem-loop hairpin structure showing MFE values of 40–100 kcal/mol, MFEI values higher than 0.85 and more than 10 miRNA reads anchored in the same orientation in the 5p and/or 3p arm in a two block-like pattern were considered putative pre-miRNAs. The precursor identity was determined by BLAST searches against mature miRNAs in miRBase and the NCBI database. As a result, 41 known full-length plant pre-miRNA sequences were identified along with 18 miRNAs anchored in the 3p-arm and 19 miRNAs in the 5p-arm (Figure S2). The precursors had an average length of 154 bp, a CG content of 43.45%, an MFE of −53.13 and an MFEI of −0.86 (Table S3), which were similar to the pre-miRNA characteristics in other plant species [14,35,36]. Twenty conserved pre-miRNAs did not generate a hairpin structure according to MFOLD (http://mfold.bioinfo.rpi.edu/cgi-bin/rnaform1.cgi).

At the present, there is no miRNA sequence for *J. curcas* available in miRNA databases, and there is only one recently published report regarding the identification of miRNAs through the cloning of sRNA sequences, which did not provide the precursor sequence [15]. In this context, the use of deep sequencing technologies represents a powerful large scale approach for the reliable identification of miRNAs in *J. curcas* [14,32,33,38,39]. In the present study, the deep sequencing of an sRNA library from *J. curcas* mature seeds and three mRNAseq libraries from three stages of seed development allowed for the identification of conserved and species-specific miRNAs and pre-miRNAs.

Identification of Novel miRNAs and Pre-miRNAs

In addition to conserved miRNAs, 16 sequences with characteristic hairpin-like structures were BLASTed against miRBase and NCBI databases, and no homologies with previously known plant miRNAs were found; these sequences characterize novel pre-miRNAs in plants. The identified pre-miRNAs had an average length of 162 bp and average MFE, MFEI and % CG content of −58.98, −0.96 and 41.50, respectively (Figure S3 and Table S4). Ten miRNAs were anchored in the 3p-arm and 15 miRNAs in the 5p-arm of these pre-miRNAs. The most abundant novel miRNA yielded 11,899 reads (Jcu_nMIR001), and it is the sixth most abundant miRNA in *J. curcas*, suggesting an important role in this tissue. The majority of novel miRNAs are 21 nt longer (Table S5), as was observed for conserved miRNAs. Interestingly, only one (JcumiR006) of the 46 novel miRNAs identified by Wang et al. [15] through cloning was identified in the present study (corresponding to JcuMIR0015_5p in the present study, sequence

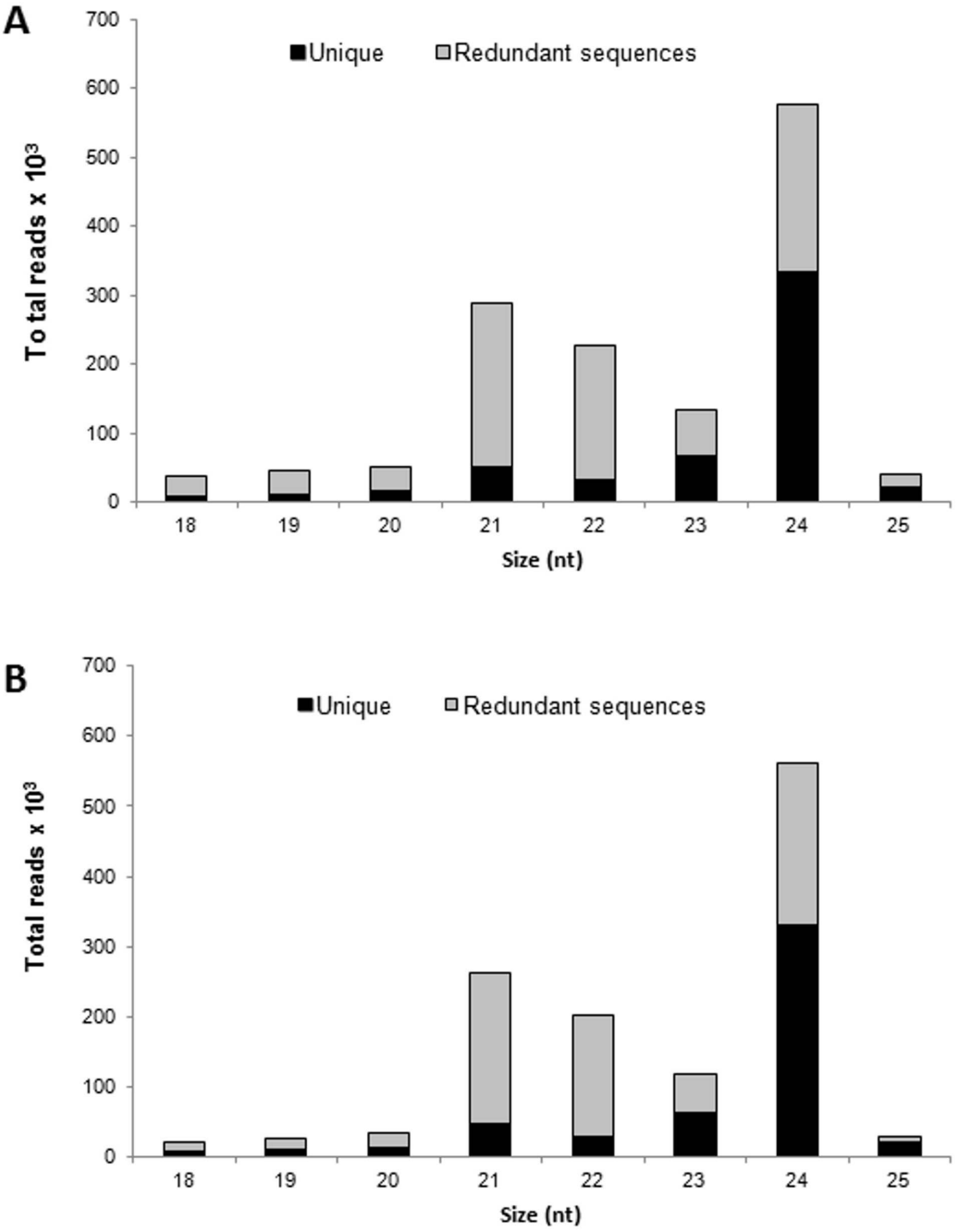

Figure 1. Total number of redundant and unique reads in the sRNA library of *J. curcas* mature seeds. (**A**) Data before filtering with non-coding RNAs and organelle RNAs. (**B**) Data after filtering with non-coding RNAs and organelle RNAs.
doi:10.1371/journal.pone.0083727.g001

GGCAUGGGCGAUAUGGGCAAGA). This difference in the identified miRNAs is most likely a result of the chosen method, as explained earlier. Similarly, members of the Jcu_MIR166 and Jcu_MIR167 families demonstrated by Vishwakarma and Jadeja [16] were also identified in the present study. However, we were unable to identify six other members from the other families in our sRNA library, suggesting that these specific microRNAs may not be expressed or may not be detectable in mature seeds.

Identification of IsomiRNAs

It was previously shown that miRNA variants, which are known as isomiRNAs, are detectable by high-throughput sequencing [40–43]. They show additional nucleotides in the 5′or 3′ terminus compared to the canonical or most abundant mature miRNAs. IsomiRNAs are considered to be a consequence of inaccuracies in Dicer pre-miRNA processing. Length heterogeneity can also arise by the exonuclease 'nibbling' of the ends, which produces a shorter template product, or by the post-transcriptional addition of one or more bases [44]. In the present study, the alignment of the sRNA library with identified *J. curcas* pre-miRNAs allowed for the

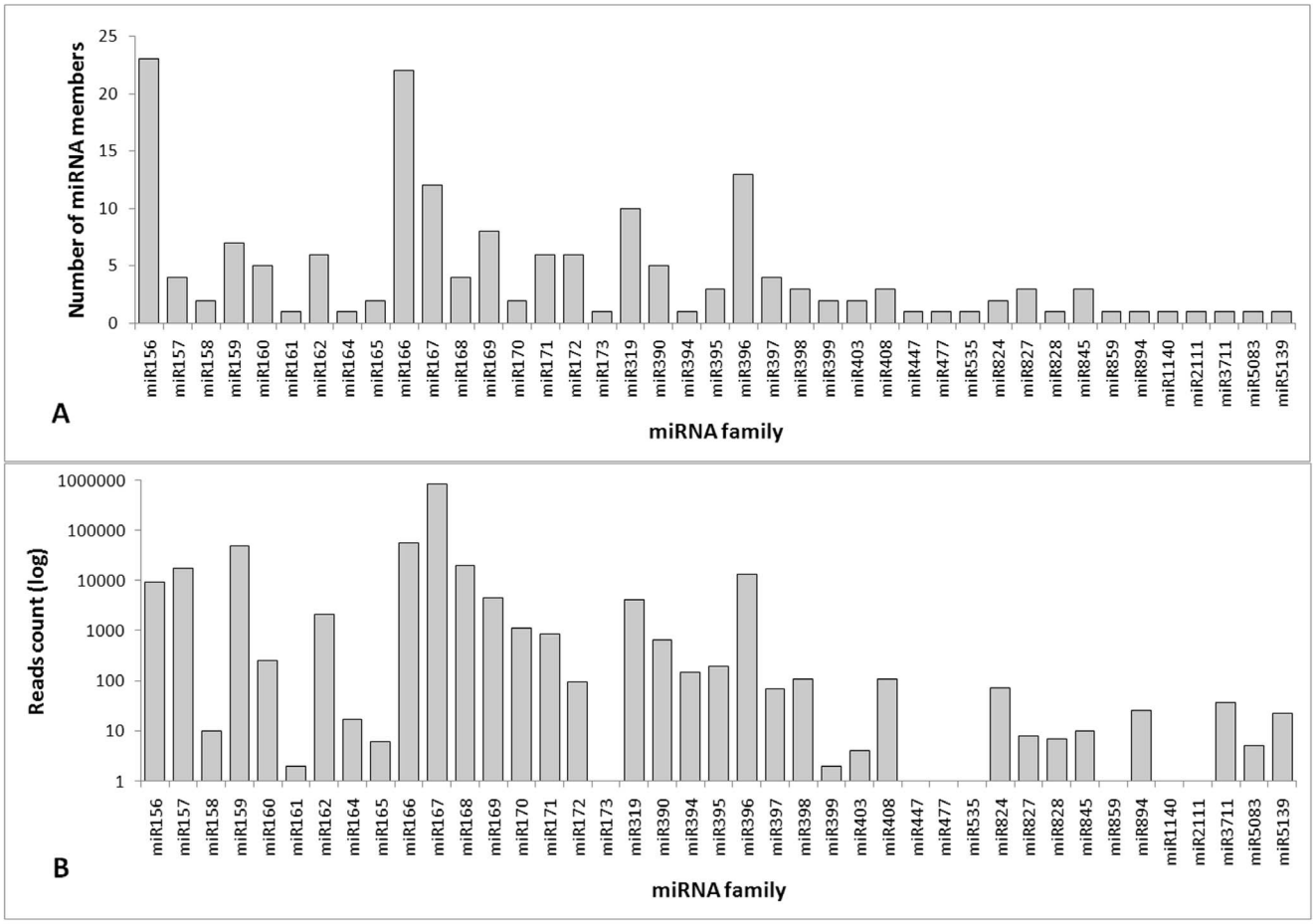

Figure 2. Known miRNA families identified in mature *J. curcas* seeds. (**A**) The total number of miRNA members (isomiRNAs) from each miRNA family. (**B**) The number of total read counts of each miRNA family.
doi:10.1371/journal.pone.0083727.g002

estimation of the number and abundance of each isomiRNA corresponding to conserved (Table S2 and Figure S2) and novel *J. curcas* miRNA families (Figure S3 and Table S5).

IsomiRNAs were produced from both known and novel predicted pre-miRNAs of *J. curcas*. It was possible to observe that miRNA families differ significantly from each other in the number and abundance of isomiRNAs, as was observed in other studies [14,34,36]. The known pre-miRNA Jcu_MIR168 and the novel pre-miRNA Jcu_nMIR001 produced more isomiRNAs than the other pre-miRNAs. Unexpectedly, a variant of the novel Jcu_nMIR001 showed more than 10,000 reads because species-specific miRNAs usually present low levels of expression compared to the conserved miRNAs. The predicted targets of Jcu_nMIR001 miRNA are ribosomal proteins, which could explain the high abundance of this miRNA. It is also interesting to note that in some conserved miRNA families, the most abundant miRNA was not the canonical miRNA described for other species. This wide variation suggests that the same miRNA family is involved in divergent functions and may be necessary at different levels, according to the species, timing, tissue and/or other situations, such as environmental conditions and stresses. It has been shown that isomiRNAs can be expressed in a cell-specific manner, and numerous recent studies suggest that at least some isomiRNAs may affect target selection, miRNA stability, or loading into the RNA-induced silencing complex (RISC) [44].

Abundance of *J. curcas* Pre-miRNAs during Seed Development

Because pre-miRNAs are processed as mRNAs species, *in silico* approaches can be used as a reliable tool for estimating the abundance of pre-miRNAs. Although this analysis does not directly predict the abundance of mature miRNA or the isomiRNAs, the number of precursor sequences present in seed developmental stages can provide information about the overall variation of the miRNA set generated from each precursor. The abundance analysis revealed that the majority of pre-miRNAs were present at similar levels in all libraries, suggesting that they are necessary during all seed development processes (Tables S3 and S4). Nevertheless, the abundance of some pre-miRNAs varied from one library to another, suggesting that the miRNAs from these precursors may be associated with the mRNA silencing involved in specific seed stage process, such as growth or maturation. Among the pre-miRNAs that were differentially represented, Jcu_MIR390 was only detected in L3 seeds; Jcu_MIR169a was only detected in L2 seeds; Jcu_MIR167a, Jcu_MIR172 and Jcu_MIR845 were only detected in L1 seeds (Table S3); Jcu_MIR156d, Jcu_MIR167b and Jcu_MIR319 were better represented in the L2 and L3 libraries; and Jcu_MIR390 was better represented in L3. Among the novel *J. curcas* miRNAs, Jcu_nMIR001 was better represented at the beginning of seed development, and Jcu_nMIR003, Jcu_nMIR007 and Jcu_n-

MIR016 were more represented in the intermediate seed stage (Table S4). It is also important to note that some novel pre-miRNAs were highly abundant, such as Jcu_nMIR001, which was the most expressed precursor during immature and mature seed stages, and Jcu_nMIR003, which was the most expressed in the intermediate stage, indicating that they play an important role in this plant. In summary, the prediction of pre-miRNA abundance among libraries increased the understanding and implications of post-transcriptional regulation in *J. curcas* seeds.

siRNA in *J. curcas* Seeds

In contrast to the biogenesis and action of miRNA, siRNAs are 24-nt long sequences produced by double-stranded RNAs that can act at the transcriptional level through DNA methylation and histone modification and at the post-transcriptional level through the regulation of gene expression [45]. Several siRNAs have been recognized to play important roles in plant stress tolerance [46,47]. Because of the large abundance of 24-nt sequences in the *J. curcas* sRNA library, we investigated the presence of siRNAs. The sRNAs that were 24 nt were matched against contigs assembled from L1–L2–L3 libraries. Putative contigs with typical sRNA distribution patterns along the matching sequences were further subjected to annotation (see methods). As a result, 42 siRNA precursors were identified and annotated (Table S6). This analysis indicates that siRNAs play a role in nuclear functions and in transport, catalytic processes and disease resistance. As expected, transposons and retroelements were relatively abundant among siRNA precursors, supporting their mechanism of action in guide chromatin-based events and resulting in transcriptional silencing [48]. It was reported that siRNA precursors can also be formed by cellular RNA-dependent RNA polymerase activity (RdRp) [49]. In fact, most *J. curcas* siRNA precursors were annotated as RdRps. Out of these precursors, the majority are associated with nuclear functions and play roles in transport, catalytic processes and disease resistance.

Prediction of *J. curcas* miRNA Targets

Because the roles of miRNAs during plant development and in species-specific adaptation processes are executed through the cleavage or translation repression of target genes [6], miRNA target prediction is critical for gaining insight into the regulatory functions of miRNAs. In this study, the putative target genes of *J. curcas* miRNAs were predicted by using the web-based computer server psRNATarget, based on perfect or near perfect complementarity between miRNAs and their targets. The most abundant mature miRNAs from each family (isomiRNAs were not considered) were aligned with a set of unigenes generated from assembled *J. curcas* contigs from all seed development mRNA-seq libraries (L1–L2–L3). A total of 57 sequences were predicted as potential targets of 28 known plant miRNAs and 12 novel miRNAs, with an average of 1.8 targets per miRNA (Table S7).

All of the identified targets were analyzed by using BLASTX against protein databases, followed by a GO analysis to evaluate their putative functions. According to the categorized GO annotation, 109 genes are involved in cellular components, with the majority of conserved and novel miRNA targets localized in intracellular membrane-bounded organelles. In the molecular functions category, 107 genes participate in catalytic or signaling transduction activities and binding activities with proteins and nucleic acids (Figure 3). With respect to biological processes, 216 genes primarily participate in stimulus responses and different cellular and metabolic processes, suggesting that the novel and conserved *J. curcas* miRNAs are involved in a broad range of physiological functions. These functions include participation in plant growth and development (pentatricopeptide repeat-containing protein, auxin response factor 10, seed maturation protein, etc.), lipid metabolism (phosphatidylserine decarboxylase and glycerophosphoryl diester), nutrient/cellular transport (amino acid transporter, high affinity nitrate transporter, metal transporter nramp6-like, etc.), and, primarily, in defense (tir-nbs-lrr resistance protein, cc-nbs-lrr resistance protein, cytosolic class I small heat shock protein partial, proline synthetase associated, dehydration responsive element binding proteins, etc.) (Table S7). The miRNAs targets were differentially represented during seed development, suggesting that they are regulated according to seed needs. For example, caffeoyl-o-methyltransferase and seed maturation protein are the most represented targets at the end of seed development, and subtilisin-like protease-like and auxin response factor 10 are better represented at the beginning of seed development than the other targets.

The auxin response factor (ARF) is a plant-specific family of DNA binding proteins involved in hormone signal transduction [50,51]. The ARF gene and the F-box family proteins, also previously described in relation to auxin signaling [52], were predicted targets of conserved *J. curcas* miRNAs and were highly abundant in immature seeds. These genes, as well as some other predicted targets, such as proteins associated with nucleotide synthesis, ribosomal proteins, metal transporters and membrane proteins, may play a role in seed growth and formation. Another important gene targeted by Jcu_MIR156, Jcu_MIR168, Jcu_-MIR403, Jcu_MIR472 and the novel Jcu_nMIR005 and Jcu_nMIR009 is the pentatricopeptide repeat gene (PPR), which participates in the regulation of gene expression and was present at considerable levels in all seed developmental stages, especially immature *J. curcas* seeds (Table S7). PPR belongs to a large gene family implicated in post-transcriptional processes, such as splicing, editing, processing and translation, in specific organelles, such as mitochondria and chloroplasts [53].

Several predicted targets of *J. curcas* miRNAs are involved in abiotic stresses, including genes associated with proline and phenylpropanoid synthesis, hormones and responses to dehydration and high temperatures. The dehydration-responsive element/C-repeat (DRE/CRT) was predicted as the target of Jcu_MIR156 and Jcu_MIR168. DRE/CRT has been identified as a cis-acting element involved in one of the ABA independent regulatory systems of abiotic stress response. In Arabidopsis, MIR156 and MIR168 were described as dehydration stress-responsive miRNAs [54]. An analysis of *J. curcas* DREB gene expression was performed by Tang et al. [55], and the expression was induced by cold, salt and drought stresses. Although only a few reads corresponding to this gene were detected in the present study, it could be interesting to investigate the role of miRNAs in the regulation of DREB expression in other tissues and during abiotic stresses, especially in *J. curcas* plants that are well-known for their tolerance to drought stress [56].

In addition to abiotic stress, miRNAs from *J. curcas* seeds may also participate in biotic stresses. The recognition of an invading microbial pathogen is often followed by the synthesis of specific plant disease resistance proteins (R) [57] that possess activities that can lead to plant cell death through the familiar hypersensitive response (HR), a characteristic feature of many plant defense mechanisms [58]. Genes encoding R proteins (cc-nbs-lrr resistance protein, disease resistance rpp13-like protein 1-like and tir-nbs-lrr resistance protein) and an HR protein were the predicted targets of Jcu_MIR159 and Jcu_MIR472 and were highly abundant at the end of seed development (Table S7), most likely as a mechanism to protect the mature seed against pathogens, thereby preventing unsuccessfully germination.

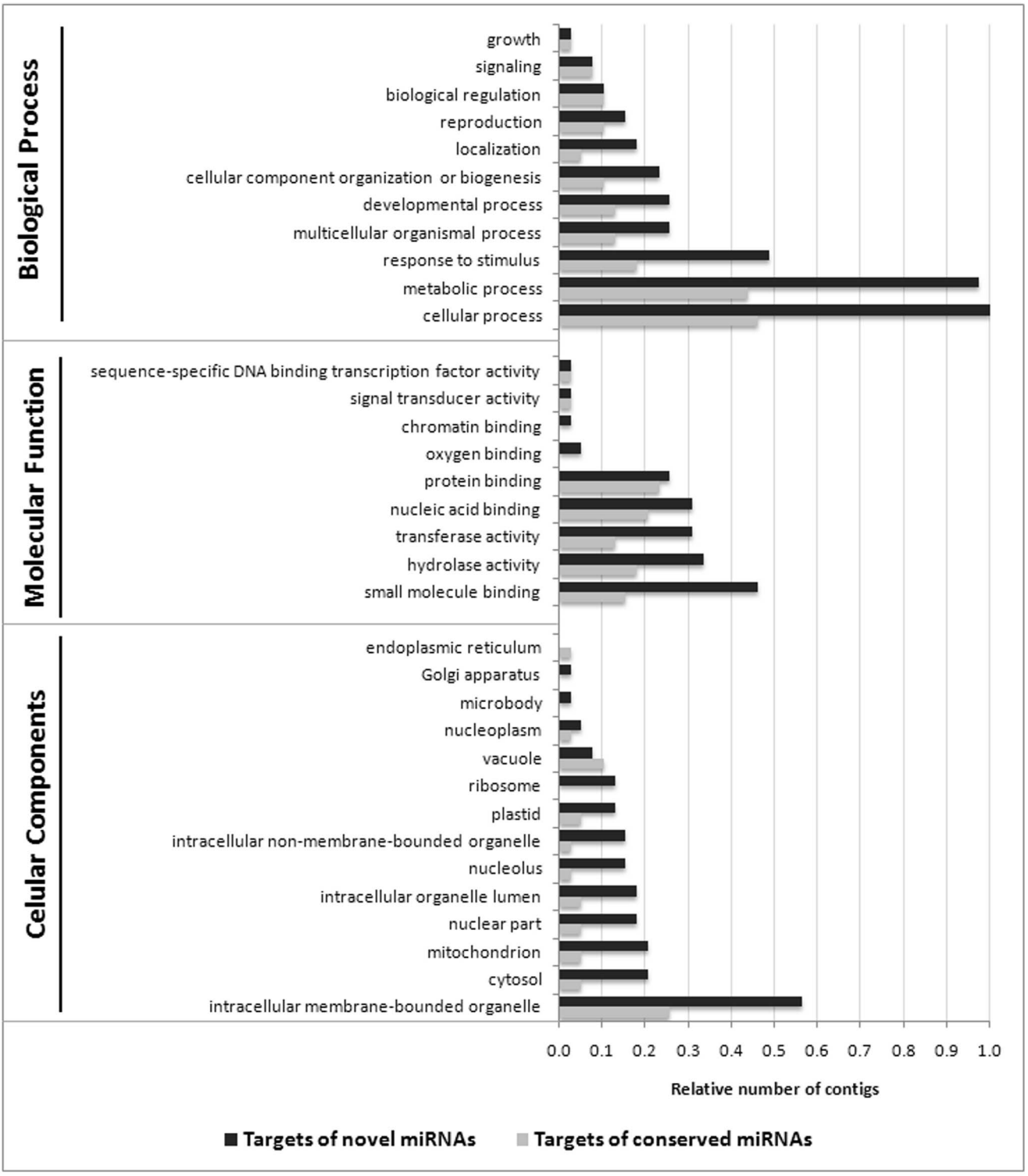

Figure 3. Targets of the miRNAs identified in mature seeds of *J. curcas*. The percentage (%) of contigs for each Gene Ontology (GO) term is relative to the total number of contigs from each gene category.
doi:10.1371/journal.pone.0083727.g003

Prediction of *J. curcas* miRNA Targets Involved in Lipid Metabolism

Jatropha has emerged as a promising biodiesel crop, but increasing the oil content and improving the oil quality are still challenging [2–4]. To investigate the involvement of isomiRNAs in regulating the lipid metabolism of *J. curcas*, the putative target genes of isomiRNAs with more than 10 reads were searched. This approach allowed for the identification of 12 miRNA targets related to lipid metabolic pathways (Table S8). This is the first report of miRNA associations with genes from the lipid

metabolism in *J. curcas*. When estimating the sequence abundance of the target contigs, only hydroxysteroid 11-beta-dehydrogenase 1-like protein was better represented at the end of seed development, and phosphatidylserine decarboxylase, diphosphomevalonate decarboxylase, 1-phosphatidylinositol-bisphosphate, lipid binding, phosphatidylserine synthase 2 and cyclopropane-fatty-acyl-phospholipid synthase were more represented at the beginning of seed formation.

LPAT, or 1-acyl-sn-glycerol-3-phosphate acyltransferase, was the predicted target of Jcu_MIR001 and Jcu_MIR007 isomiRNA. This enzyme provides the key intermediate in membrane phospholipid and storage lipid biosynthesis in developing seeds [59]. Through the analysis of miRNA target abundance, it was possible to observe that LPAT was more represented in intermediate seeds. It was proposed that increasing the LPAT expression in seeds leads to a greater flux of intermediates through the Kennedy pathway and enhanced triacylglycerol accumulation [43]. Glycerophosphoryl diester (predicted target of Jcu_MIR001) and glycerol-3-phosphate dehydrogenase (predicted target of Jcu_MIR403) provide glycerol-3-phosphate for TAG assembly in the Kennedy pathway, and they were also more represented in intermediate seeds. These results confirm the general observation that storage lipid biosynthesis usually occurred at the middle-late stage during seed development [60] and suggest that the seed already reached maturity by the intermediate stage. Interestingly, a seed maturation protein target of Jcu_MIR472 was better represented in the intermediate stage, corroborating this finding. The same LPAT and GPDH expression pattern during seed development was observed by Xu et al. [61] and Guo et al. [43] by using real time PCR, which confirmed these results. In the study by Xu et al. [61], the authors verified that the gene encoding fatty acid desaturase exhibited low expression in all seed developmental stages, as was observed in the present study, suggesting that the transcript accumulation of this gene is not necessarily linked to triacylglycerol biosynthesis in developing *Jatropha* seeds.

Conclusions

There is some uncertainty about the genetic contribution to divergent phenotypes and the response to growth conditions and stress in *Jatropha* genotypes because there are several reports indicating low genetic diversity among these plants. Epigenetic polymorphisms were suggested to be involved [62] but may not explain all of the observed the variation. The present study identified the *J. curcas* miRNAs that are involved in a wide range of physiological functions and therefore may also contribute to this phenotypic variation. The identification of a large set of miRNAs and their targets as well as siRNAs in *J. curcas* seeds contributes to the elucidation of complex miRNA-mediated regulatory systems, which control seed development and other physiological processes. Several *Jatropha* miRNAs were predicted to regulate genes associated with lipid metabolism; these miRNAs are promising candidates for improving the yield and quality of *J. curcas* seed oil. However, further studies are necessary to search for more novel miRNAs and to validate their target by expression analysis during seed development and also under specific environmental and physiological conditions.

Supporting Information

Figure S1 Flow chart of the methodology adopted to identify *J. curcas* miRNAs.
(PDF)

Figure S2 Predicted secondary structures of known miRNA precursors in *J. curcas*. Locations and expressions of small RNAs mapped onto these precursors are presented. Read sequences corresponding to miRNA candidates, which are located in the 5p and 3p arms and labeled in red and purple, respectively. Values on the left side of the miRNA sequences represent the miRNA length (Jn) and read counts (x n) in the mature seed library.
(PDF)

Figure S3 Predicted secondary structures of novel miRNA precursors in *J. curcas*. The locations and the expression of small RNAs mapped onto these precursors are shown here. The sequences of miRNA candidates located in the 5p and 3p arms are labeled in red and purple, respectively. Values on the left side of the miRNA sequence represent miRNA length (Jn) and read counts (x n) in the mature seed library.
(PDF)

Table S1 Abundance of small RNAs from each size according to the number of reads.
(XLSX)

Table S2 Known miRNAs identified in the *J. curcas* sRNA library. This identification was performed by selecting perfect homologies to plant-conserved miRNAs, as deposited in the miRBase database (release 19, August 2012).
(XLSX)

Table S3 Putative known miRNA precursors in *J. curcas* and their abundance during seed development.
(XLSX)

Table S4 Putative new *J. curcas* miRNA precursors and their abundance during seed development.
(XLSX)

Table S5 Novel miRNAs identified in the sRNA library of *J. curcas*.
(XLSX)

Table S6 Putative precursors of siRNA in *J. curcas*.
(XLSX)

Table S7 Predicted targets of known and new miRNAs in mature *J. curcas* seeds. The abundance of these targets during seed development is shown here.
(XLSX)

Table S8 Predicted targets of known and new isomiRNAs related to the lipid metabolism of *J. curcas*. The abundance of targets during seed development is shown here.
(XLSX)

Author Contributions

Conceived and designed the experiments: VG SDAS RM. Performed the experiments: VG. Analyzed the data: VG FG LFVO GLM APK. Contributed reagents/materials/analysis tools: LFVO GLM SDAS MMP RM. Wrote the paper: VG FG LFVO GLM APK RM.

References

1. Carels N (2009) *Jatropha curcas*: A Review. Adv Bot Res 50: 39–86.
2. Divakara BN, Upadhyaya HD, Wani SP, Laxmipathi CL (2010) Biology and genetic improvement of *Jatropha curcas* L. Appl Energ 87: 732–742.
3. Berchmans HJ, Hirata S (2008) Biodiesel production from crude *Jatropha curcas* L. seed oil with a high content of free fatty acids. Bioresour Technol 99: 1716–1721.

4. Becker K, Makkar HPS (2008) *Jatropha curcas*: A potential source for tomorrow's oil and biodiesel. Lipid Tech 20: 104–107.
5. King AJ, He W, Cuevas A, Freudenberger M, Ramiaramanana D, et al. (2009) Potential of *Jatropha curcas* as a source of renewable oil and animal feed. J. Exp. Bot. 60: 2897–2905.
6. Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9: 102–114.
7. Lee Y, Kim M, Han J, Yeom KH, Lee S, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23: 60–4051.
8. Bartel DP (2009) MicroRNAs: Target recognition and regulatory functions. Cell 136: 215–233.
9. Voinnet O (2009) Origin, biogenesis and activity of plant microRNAs. Cell 136: 669–687.
10. Huntzinger E, Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet 12: 99–110.
11. Chen H, Li Z, Xiong L (2012) A plant microRNA regulates the adaptation of roots to drought stress. FEBS Letters 586: 1742–1747.
12. Schommer C, Bresso EG, Spinelli SV, Palatnik JF (2012) MicroRNAs in plant development and stress responses. Signaling and Communication in Plants 15: 29–47.
13. Chi X, Yang Q, Chen X, Wang J, Pan L, et al. (2011) Identification and characterization of microRNAs from peanut (*Arachis hypogaea* L.) by high-throughput sequencing. PLoS ONE 6: e27530.
14. Körbes AP, Machado RD, Guzman F, Almerão MP, de Oliveira LFV, et al. (2012) Identifying conserved and novel microRNAs in developing seeds of *Brassica napus* using deep sequencing. PLoS ONE 7: e50663.
15. Wang CM, Liu P, Sun F, Li L, Liu P, et al. (2012) Isolation and identification of miRNAs in *Jatropha curcas*. Int J Biol Sci 8: 418–429.
16. Vishwakarma NP, Jadeja VJ (2013) Identification of miRNA encoded by *Jatropha curcas* from EST and GSS. Plant Signal Behav 8: 2, e23152.
17. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864.
18. Jühling F, Mör M, Hartmann RK, Sprinz M, Stadler PF, et al (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res 37: 159–162.
19. Pruesse E, Quast C, Knitte K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196.
20. He G, Elling AA, Deng XW (2011) The Epigenome and Plant Development. Annu Rev Plant Biol 62: 411–435.
21. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.
22. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.
23. Mathews DH (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 10: 1178–90.
24. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, et al. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651–652.
25. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868–877.
26. Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 39: 9–155.
27. Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008: 1–12.
28. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25.
29. Robinson MD, McCarthy DJ, Smyth GK (2010) EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–40.
30. Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res 7: 986–995.
31. Molina LG, Cordenonsi FG, Morais GL, de Oliveira LFV, Carvalho JB, et al. (2012) Metatranscriptomic analysis of small RNAs present in soybean deep sequencing libraries. Genet Mol Biol 35: 292–303.
32. Hsieh LC, Lin SI, Shih AC, Chen JW, Lin WY, et al. (2009) Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. Plant Physiol 151: 2120–2132.
33. Schreiber AW, Shi BJ, Huang CY, Langridge P, Baumann U (2011) Discovery of barley miRNAs through deep sequencing of short reads. BMC Genomics 12: 129.
34. Song QX, Liu YF, Hu XY, Zhang WK, Ma B, et al. (2011) Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. BMC Plant Biology 11: 5.
35. Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA (2006) Conservation and divergence of plant microRNA genes. The Plant Journal 46: 243–259.
36. Sunkar R, Agadeeswaran G (2008) In silico identification of conserved microRNAs in large number of diverse plant species. BMC Plant Biol 8: 1–13.
37. Guzman F, Almerão MP, Körbes AP, Loss-Morais G, Margis R (2012) Identification of microRNAs from *Eugenia uniflora* by high-throughput sequencing and bioinformatics analysis PLoS ONE 7: e49811.
38. Yian Z, Li C, Han X, Shen F (2008) Identification of conserved microRNAs and their target genes in tomato (*Lycopersicon esculentum*). Gene 414: 60–66.
39. Moxon S, Schwach F, Maclean D, Dalmay T, Studholme DJ, et al. (2008) A tool kit for analysing large-scale plant small RNA datasets. Bioinformatics 24: 2252–2253.
40. Kulcheski FR, de Oliveira LF, Molina LG, Almerão MP, Rodrigues FA, et al. (2011) Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC Genomics 12: 307.
41. Lelandais-Brière C, Naya L, Sallet E, Calenge F, Frugier F, et al. (2009) Genome-wide *Medicago truncatula* small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules. Plant Cell 21: 2780–2796.
42. Ebhardt A, Fedynak A, Fahlman RP (2010) Naturally occurring variations in sequence length creates microRNA isoforms that differ in argonaute effector complex specificity. Silence 1: 12.
43. Gu K, Yi C, Tian D, Sangha JS, Hong Y, et al. (2012) Expression of fatty acid and lipid biosynthetic genes in developing endosperm of *Jatropha curcas*. Biotechnology for Biofuels 5: 1–15.
44. Yu X, Wang H, Lu Y, de Ruiter M, Cariaso M, et al. (2012) Identification of conserved and novel microRNAs that are responsive to heat stress in *Brassica rapa*. J. Exp. Bot 63: 1025–38.
45. Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. Cell 136: 642–655.
46. Dunoyer P, Brosnan CA, Schott G, Wang Y, Jay F, et al. (2010) An endogenous, systemic RNAi pathway in plants. The EMBO Journal 29: 1699–1712.
47. Khraiwesh B, Zhu JK, Zhu J (2012) Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. Biochimi Biophys Acta - Gene Regulatory Mechanisms 1819: 137–148.
48. He S, Liu C, Skogerbo G, Zhao H, Wang J, et al. (2008) NONCODE v2.0: decoding the non-coding. Nucleic Acids Res 36: 170–172.
49. Ahlquist P (2002) RNA-Dependent RNA polymerases, viruses, and RNA silencing. Science 296: 1270–1273.
50. Yang JH, Han SJ, Yoon EK, Lee WS (2006) Evidence of an auxin signal pathway, microRNA167-ARF8-GH3, and its response to exogenous auxin in cultured rice cells. Nucleic Acids Res 34: 1892–1899.
51. Wu MF, Tian Q, Reed JW (2006) Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction. Development 133: 4211–4218.
52. Kipreos ET, Pagano M (2000) The F-box protein family. Genome Biol 1: 1–7.
53. Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci 13: 663–670.
54. Liu H, Tian X, Li Y, Wu C, Zheng C (2008) Microarray-based analysis of stress-regulated microRNAs in *Arabidopsis thaliana*. RNA 14: 836–843.
55. Tang M, Liu X, Deng H, Shen S (2011) Over-expression of JcDREB, a putative AP2/EREBP domain-containing transcription factor gene in woody biodiesel plant *Jatropha curcas*, enhances salt and freezing tolerance in transgenic Arabidopsis thaliana. Plant Sci 181: 623–631.
56. Wang WG, Li R, Liu B, Li L, Wang SH, et al. (2011) Effects of low nitrogen and drought stresses on proline synthesis of *Jatropha curcas* seedling. Acta Physiol Plant 33: 1591–1595.
57. Belkhadir Y, Subramaniam R, Dangl JL (2004) Plant disease resistance protein signaling: NBS–LRR proteins and their partners. Curr Opin Plant Biol 7: 391–399.
58. Greenberg JT, Yao N (2004) The role and regulation of programmed cell death in plant-pathogen interactions. Cellular Microbiol 6: 201–211.
59. Maisonneuve S, Bessoule JJ, Lessire R, Delseny M, Roscoe TJ (2010) Expression of rapeseed microsomal lysophosphatidic acid acyltransferase isozymes enhances seed oil content in Arabidopsis. Plant Physiol 152: 670–684.
60. Hills MJ (2004) Control of storage-product synthesis in seeds. Curr Opin Plant Biol 7: 302–8.
61. Xu R, Wanga R, Liu A (2011) Expression profiles of genes involved in fatty acid and triacylglycerol synthesis in developing seeds of Jatropha (*Jatropha curcas* L.). Biomass and bioenergy 35: 1683–1692.
62. Yi C, Zhang S, Liu X, Bui HT, Hong Y (2010) Does epigenetic polymorphism contribute to phenotypic variances in *Jatropha curcas* L.? BMC Plant Biology 10: 259.